Technical Report ITL-98-3
September 1998

US Army Corps of Engineers
Waterways Experiment Station

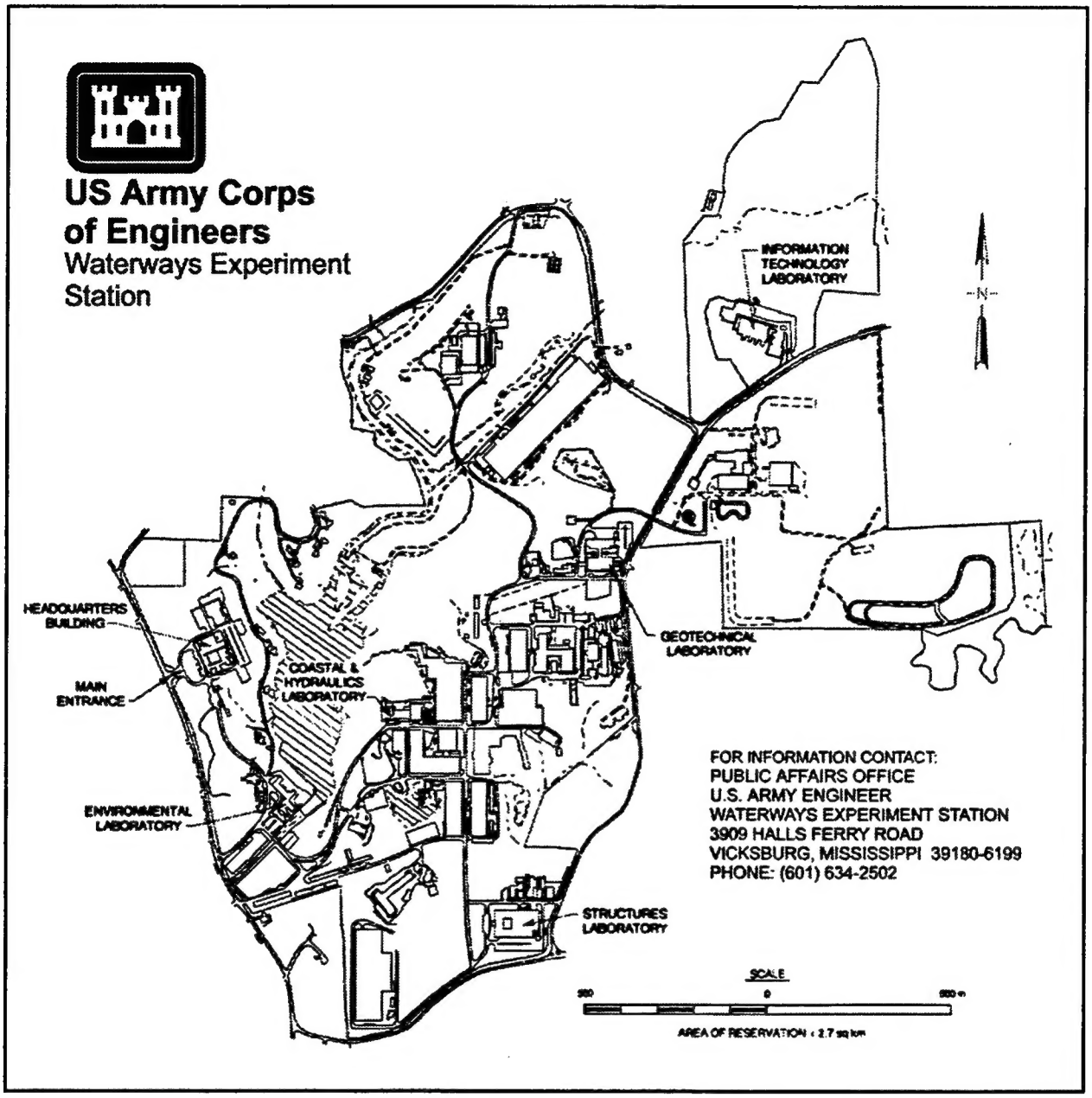
Displayless Interface Access to Spatial Data: Effects on Speaker Prosodics

by Julia A. Baca

The contents of this report are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products.

The findings of this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

PRINTED ON RECYCLED PAPER



Waterways Experiment Station Cataloging-in-Publication Data

Baca, Julia A.

Displayless interface access to spatial data : effects on speaker prosodics / by Julia A. Baca.

249 p. : ill. ; 28 cm. -- (Technical report ; ITL-98-3)

Includes bibliographical references.

1. Human-computer interaction. 2. User interfaces (Computer systems) 3. Natural language processing (Computer science) 4. Computer interfaces. I. United States. Army. Corps of Engineers. II. U.S. Army Engineer Waterways Experiment Station. III. Information Technology Laboratory (U.S. Army Engineer Waterways Experiment Station) IV. Title. V. Title: Effects on speaker prosodics. VI. Series: Technical report (U.S. Army Engineer Waterways Experiment Station) ; ITL-98-3.

TA7 W34 no.ITL-98-3

PREFACE

This report documents research in the use of displayless interface technology. Specifically, the research investigates the impact of the use of such technology on the nonverbal or prosodic aspects of user speech. This research was conducted in the Department of Computer Science, Mississippi State University, by Ms. Julia A. Baca in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the direction of Dr. Julia E. Hodges. The work was sponsored by the U.S. Army Engineer Waterways Experiment Station (WES), Information Technology Laboratory (ITL), under the direction of Dr. Windell F. Ingram, Chief, Computer Science Division, and Dr. N. Radhakrishnan, Director, ITL.

At the time of publication of this report, Director of WES was Dr. Robert W. Whalin. Commander was Colonel Robin R. Cababa, EN.

TABLE OF CONTENTS

Page

PREFACE ii

LIST OF TABLES viii

LIST OF FIGURES xvii

CHAPTER

I. INTRODUCTION 1

II. LITERATURE REVIEW 5

Cognition: Verbal versus Spatial 5
Neurophysiology: Left Brain versus Right Brain 5
Psychological Studies of Memory and Recall 8
Education and Learning: Cognitive Styles 11
Human-Computer Interaction 12
Cognitive and Perceptual Differences for Individuals with Sight Loss . 14
Spatial and Motor Development 15
Cognitive Development 16
Perceptual Abilities 17
Adventitious Versus Congenital Sight Loss 19
Summary 20
GUI Access 21
Early Obstacles to GUI Access 21
Open Issues for GUI Access 22
Current GUI Access Research 23
Non-Speech Auditory Cues 23
Tactile Interaction 25
Displayless Interfaces 26
Applications Appropriate for Speech Interfaces 26

CHAPTER Page

Verbal Cognition 29
Auditory Overload 30
Prosodics 31
Summary 32
Speaker Prosodics 33
Language-Independent Prosodic Features 34
Pauses 34
F0 Features 35
Declination Tendency 35
Normal Frequency Range and Control 35
Rising versus Falling F0 Movements 36
Durational Features 37
Final Lengthening 37
Other Durational Lengthening 37
Prosodic Pattern Detection Algorithms 37
Prosodic Phrase Detection 38
Tune Recognition 40
Prominence Detection 40
Analysis 44
Acoustical Correlates of Emotion 45
Physiological Effects 45
Studies of Simulated Emotion 46
Studies of Laboratory-Induced Stress 49
Real-Life Studies 52
Summary 53

III. EXPERIMENTAL DESIGN 56

Experimental Testing 56
General Approach 56
Experimental Treatments 57
Spatial Complexity of Tasks 57
Subject Preparation for Testing 61
Subjects 62
General Criteria 62
Criteria for Subjects with Visual Impairments 63
Criteria for Sighted Subjects 64
Data Analysis 65
User Data: Prosodics 65
Prosodic Variables 65

CHAPTER Page

Statistical Analysis of Prosodic Variables 65
System Data: Speech Recognition Errors 66
Categories of Recognition Errors 66
Experimental Measurements of Recognition Errors 68
Statistical Analysis of Recognition Errors 70
Scope of Study 71

IV. EXPERIMENTAL PROCEDURES

Speech-based Prototype Navigational System 73
Overview 73
GIS Database Module 76
Design 76
Data Design 76
Application Design 78
Implementation 79
Natural Language Module 81
Design 81
Implementation 87
Parser/Translator 88
Phrase Generator 92
Speech Input Module 94
Design 94
Implementation 95
Usability Testing 96
Speech Output Module 102
Design 102
Implementation 103
Symmetry of Commands 104
Interruptibility. 105
Multimodal Interface 106
Design 106
Graphical Interface 106
Tactile Interface 108
Implementation 109
Graphical Interface 109
Tactile Interface 110
Prosodic Labelling Method 110
Requirements 110
Transcription System 111

CHAPTER Page

Tonal Tier 112
Break Index Tier 113
Orthographic Tier 114
Miscellaneous Tier 115

V. EXPERIMENTAL RESULTS 116

Testing Conditions 116
Analyses of Data from Multimodal versus Displayless Sessions 118
Subjects with Congenital Vision Loss 118
Pauses 118
Intonational Boundary Tones 120
Durational Features 121
Recognition Errors 124
Subjects with Adventitious Vision Loss 125
Pauses 126
F0 and Intonational Features 127
Recognition Errors 130
Sighted Subjects 131
Pauses 132
F0 and Intonational Features 133
Durational Features 135
Recognition Errors 136
Summary 137
Task-level Analyses of Data from Displayless versus Multimodal Sessions 140
Subjects with Congenital Vision Loss 141
Pauses 141
Intonational Features 144
Durational Features 145
Recognition Errors 149
Subjects with Adventitious Vision Loss 151
Pauses 151
Intonational Features 153
Durational Features 156
Recognition Errors 160
Sighted Subjects 161
Pauses 161
Intonational Features 164
Durational Features 167

CHAPTER Page

Recognition Errors 170
Summary of Task-level Results 171
Subjects with Congenital Vision Loss 172
Subjects with Adventitious Vision Loss 175
Sighted Subjects 177
Interpretation of Results 180
Human Factors Issues 180
Recognition Errors and Prosodics 181
Prosodics and Cognitive Load 182
Cognitive Load 183
Limitations of Study 186
Sighted Subjects Categorized by Cognitive Preference 189
Relevance of Results for Prosodic Pattern Detection Algorithms 193

VI. CONCLUSIONS 195

REFERENCES 199

APPENDIX

A. SUBJECT INSTRUCTIONS FOR EXPERIMENT 213

B. SUBJECT QUESTIONNAIRES 216

C. MAP FOR MULTIMODAL SESSION 228

LIST OF TABLES

TABLE Page

1. Matched-pair T Test Results for Number of Pauses per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 119

2. Matched-pair T Test Results for Average Pause Length per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 120

3. Matched-pair T Test Results for Number of Occurrences of Boundary Tones per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 121

4. Matched-pair T Test Results for Durational Features of Utterances Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 122

5. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 123

6. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 123

7. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 124

8. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss 125

TABLE Page

9. Matched-pair T Test Results for Number of Pauses per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 126

10. Matched-pair T Test Results for Average Pause Length per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 127

11. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 128

12. Matched-pair T Test Results for Number of Occurrences of Boundary Tones per Utterance Spoken in Displayless vs. Multimodal Sessions By Subjects with Adventitious Vision Loss 128

13. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 129

14. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 129

15. Matched-pair T Test Results for Durational Features of Utterances Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 130

16. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss 131

17. Matched-pair T Test Results for Number of Pauses per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects . . 132

18. Matched-pair T Test Results for Average Length of Pauses per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects . . 133

TABLE Page

19. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 134

20. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 134

21. Matched-pair T Test Results for Durational Features of Utterances Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 135

22. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 136

23. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 136

24. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects 137

25. Summary of Significantly Differing Variables in Overall Sessions 139

26. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 142

27. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 142

28. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 143

29. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 143

TABLE Page

30. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 144

31. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 145

32. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 146

33. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 146

34. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 147

35. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 147

36. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 148

37. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 148

38. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 148

39. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 149

TABLE Page

40. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 150

41. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2 150

42. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 152

43. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 152

44. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 153

45. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 153

46. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 154

47. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 154

48. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 155

49. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 156

TABLE Page

50. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 157

51. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 158

52. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 159

53. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 159

54. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 159

55. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 160

56. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 160

57. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 161

58. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 162

59. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 162

TABLE Page

60. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 163

61. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 163

62. Matched-pair T Test Results for Minimum and Maximum F0 Values per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 164

63. Matched-pair T Test Results for Minimum and Maximum F0 Values per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 165

64. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 166

65. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 166

66. Matched-pair T Test Results for Durational Features of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 167

67. Matched-pair T Test Results for Durational Features of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 168

68. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 168

69. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 169

TABLE Page

70. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 169

71. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 169

72. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1 170

73. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2 171

74. Matched-pair T Test Results for Speaking Rate per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 172

75. Matched-pair T Test Results for Substitution and Insertion Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1 174

76. Matched-pair T Test Results for Minimum F0 Values Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1 176

77. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2 177

78. Summary of Task-level Results for Variables Differing Significantly in Overall Sessions 179

79. Matched-pair T Test Results for Number of Hesitation Pauses per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 190

TABLE Page

80. Matched-pair T Test Results for Average Length of Hesitation Pauses per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 190

81. Matched-pair T Test Results for Number of "L%" Boundary Tones per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 191

82. Matched-pair T Test Results for Number of "H%" Boundary Tones per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 191

83. Matched-pair T Test Results for Duration of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 191

84. Matched-pair T Test Results for Duration of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference 192

LIST OF FIGURES

FIGURE		Page
1.	Basic Route Patterns	58
2.	Possible Street Crossings	59
3.	I-Route with Street Crossings	60
4.	Displayless Access to Prototype	74
5.	Multimodal Access to Prototype	75
6.	Detailed View of NLP Module	88
7.	Subnet Produced by Phrasal Generator	93
8.	Relationship Among Cognitive Load, Prosodics, Recognition Error Rate, and User Satisfaction	185

CHAPTER I

INTRODUCTION

The emergence of the graphical user interface (GUI) has marked a turning point in modern computing environments. For sighted users, the GUI provides a more natural interaction with the computer system, allowing a direct manipulation of objects and actions within the interface (Shneiderman 1984). For users with visual impairments, however, gaining access to these interfaces has presented major challenges (Vanderheiden and Kunz 1990; Boyd et al. 1992). The use of the GUI in human-computer interaction continues only to increase with no signs of abating. This fact, coupled with the 1990 passage of the Americans with Disabilities Act (ADA), requiring employers to provide reasonable accommodations for persons with disabilities, heightens the imperative to provide users with visual impairments better access to GUI's. Though some of the initial obstacles have been addressed, many issues remain problematic.

Paradoxically, the increasing popularity of GUI's for sighted users has been followed by a recent trend, incongruous to the purpose of a GUI, the development and use of "displayless" interfaces. These interfaces offer voice access to applications in which a visual display cannot be used, such as telephone-based interactions, or applications in which the user's hands and eyes are busy with other tasks, such as

piloting aircraft. Although displayless interfaces introduce certain issues specific to spoken language understanding, similar underlying challenges apply for both displayless and GUI access interfaces. Both must address the unique problems presented by nonvisual access to data, especially data which is either inherently spatial in nature or which is presented through the use of a visual and spatial display metaphor. While some methods have been developed for dealing with the latter problem in GUI access, i.e., spatial presentation of data which may not be inherently spatial (Weber and Mynatt 1994), the problem of accessing inherently spatial information without vision remains an open issue for all users, with or without vision loss. The assumption that accomplishing such access through spoken language alone produces a cognitive burden for the individual, regardless of visual capability, serves as a basis for the proposed research.

The literature review begins by presenting a selective survey of research, originating from an amalgam of viewpoints, including psychology, neurology, education, and human-computer interaction, which supports this assumption. Particular discussion of how the assumption relates to users with visual impairments is included. Subsequent discussion details the early obstacles for these users in accessing GUI's and outlines the issues which remain unresolved. Current research in the area of nonvisual GUI's is then reviewed. This review includes discussion of why voice interaction should be considered as part of a solution for GUI access, even though it presents certain difficulties. While spoken language alone may not be the optimum means of access to spatial data, used with other input and output modalities, it offers

certain advantages; e.g., voice input frees the hands to be used for other tasks, such as accessing a tactile output device. Boyd, Boyd, and Vanderheiden (1990) advocate the development of a nonvisual multimodal interface, including voice input and tactile output as the next level of achievement in providing users with visual impairments better access to GUI's.

Therefore, an understanding of the important issues in the use of speech-based interfaces is necessary. These issues are outlined and briefly discussed in the literature review. This discussion includes an evaluation of situations in which a speech-based interface is appropriate or desirable. For certain situations a speech interface may be optimum, e.g., mobile users whose hands and eyes are busy, while for other situations, speech may be less than optimum, but necessary, e.g., accessing spatial information when vision is restricted (through an impairment or the environment). Finally, one issue in particular, speaker prosodics, is examined in depth. Prosodics encompasses the nonverbal aspects of spoken language, such as pauses and intonation, which are useful in speech recognition. Examination of this issue includes an argument which establishes a connection between the cognitive burden produced by nonvisual access to graphical data through spoken language and the manifestation of this burden in the speaker prosodics. An understanding of how this cognitive burden impacts the speaker prosodics will contribute to the development of more robust interfaces for situations in which this type of access is necessary. Insight gained from an investigation of this issue can be used, for example, to improve algorithms used in the recognition component by identifying certain prosodic patterns or variations in prosodic patterns

occurring for this type application. Also, research on "parroting," the tendency of speakers in human-computer interaction to mimic the speaking style of the computer interface (Zoltan-Ford 1991), indicates that modulating the prosodics of the computer speech may be helpful in controlling irregularities identified in the prosodics of the human speaker for these applications.

It is the hypothesis of this research that the prosodic patterns of speech produced to access visuospatial information through a displayless interface, employing only spoken language, will differ significantly from those of speech produced when the interface employs an additional output modality. Although the hypothesis is assumed to apply to all users, regardless of visual capability, experience in coping without vision can potentially affect the behavior of users with vision loss versus sighted users. In addition, the presence of visual memory distinguishes sighted users as well as users with adventitious vision loss from those with congenital vision loss. Therefore, each of these categories of subjects participated in this research, including sighted subjects, subjects with adventitious vision loss, and subjects with congenital vision loss.

Following the review of the literature given in Chapter Two, the experimental design is discussed in Chapter Three. Chapter Four describes the experimental procedures and methods in further detail. Chapter Five reports and interprets the experimental results. Finally, Chapter Six presents conclusions as well as areas for future investigation

CHAPTER II

LITERATURE REVIEW

Cognition: Verbal versus Spatial

Research conducted in many disciplines, including neurophysiology, psychology, education, and human-computer interaction has indicated fundamental differences in the way in which humans process verbal and visuospatial information. A selective survey of this research is presented in this section. Again, such research is important because it lends support to a fundamental assumption of the proposed research, the inherent difficulty of nonvisual access to visuospatial information through spoken language alone.

Neurophysiology: Left Brain versus Right Brain

Laboratory research conducted during the 1970's, using electro-encephalographic (EEG) technology, confirmed existing clinical evidence that most humans process verbal language predominantly in the left hemisphere of the brain, while visuospatial and other non-verbal cognitive processing occurs predominantly in the right hemisphere (Witelson 1976). Alpha rhythms occurring in the brain can be observed in EEGs. Alpha rhythms are an indication of the level of visual attention to stimulus; in general, fewer alpha rhythms are observed when more visual control

systems are active. They are measured by EEG's in terms of voltage, fluctuating from 8-13 Hz frequency and can be recorded for at least 85% of the population (Mulholland 1978). Research demonstrating relative left EEG activation for verbal tasks and relative right EEG activation for visuospatial tasks, including Galin and Ornstein (1962); Davidson et al. (1976); Butler and Glass (1974); Duman and Morgan (1975) is summarized by Mulholland (1978).

More recent studies have used advanced imaging techniques for visualizing the brain. Magnetic resonance imaging (MRI) is one such technique. Although not all regions of the brain associated with cognitive functioning can be visualized with this technique, certain regions can be viewed clearly. In particular, the large fiber groups such as those found in the corpus callosum can be clearly visualized. The corpus callosum, the central fiber pathway connecting the hemispheres of the brain, is considered to play a significant role in human cognition. Hines et al. (1992) cite studies by Chiarello (1980) and Sperry (1982) of individuals who have had lesions on the corpus callosum or surgical separation of the hemispheres by severing the corpus callosum which have indicated its significance in normal cognitive functioning.

A study performed by Hines et al. (1992) measured correlations of the size of specific areas of the corpus callosum to verbal fluency, visuospatial ability, and language lateralization (reliance on a single hemisphere to process verbal language). The authors cite prior MRI research suggesting gender differences in the size of specific areas of the corpus callosum (da Lacoste-Utamsing and Holloway 1982; Clarke et al. 1989) as well as psychological research reporting average differences for

men and women on visuospatial ability and verbal fluency tasks (Halpern 1987; Maccoby and Jacklin 1974). Based on this research, Hines et al. (1992) formulated a hypothesis relating the size of specific areas of the callosum to certain cognitive capabilities. Specifically, they hypothesized that the posterior callosum would correlate positively with verbal fluency and negatively with language lateralization and that the anterior callosum would correlate positively with visuospatial ability. Using a principal components analysis, the posterior callosum factor predicted positively and significantly the verbal fluency factor. The correlation between the posterior callosum factor and the language lateralization factor was negative and significant with a one-tailed test, and the correlation between the anterior callosum factor and the visuospatial ability factor was positive and significant with a one-tailed test. All three correlations observed were consistent with the hypothesis. Further details of the analysis are given in (Hines et al. 1992).

The implications of this study for possible gender differences in cognition, though interesting, are less pertinent to the discussion than the results indicating that verbal and visuospatial cognitive processes may occur in separate regions of the callosum. These results as well as the results of the EEG studies lend support to the central assumption of the proposed research, that accessing visuospatial information through spoken language alone imposes a cognitive burden on the individual, regardless of visual capability.

Psychological Studies of Memory and Recall

Evidence in the psychological literature indicates that verbal and visuospatial information may be encoded differently in both short-term and long-term memory. In fact, Klatzky (1980) notes that there is little disagreement in the literature regarding the existence of a special visual code in short-term memory (STM), although there is debate as to the exact nature of the code, whether analogue, i.e., a direct representation, or propositional, a less detailed, more abstract representation. There is greater controversy, however, regarding the existence of separate codes for visual and verbal information in long-term memory (LTM). Some of the research and argument is presented in the following paragraphs. A more thorough review is given in Klatzky (1980a, 1980b), and Ericsson and Simon (1993).

Klatzky (1980a) cites a series of experiments performed by Brooks (1968), which indicate the existence of differing visual and verbal encodings of information in STM. Subjects were given sentences to commit to memory and asked to identify, without looking at the sentence, whether each word in the sentence was a noun. Subjects took longer to verbalize the words "yes" or "no" than to point to a visual representation of the same words. This suggests that the similar encodings of the words in the sentence competed with each other; in other words, verbalizing "yes" or "no" competed with the verbal code held in memory representing the words in the sentence, while pointing to visual representations did not.

Ericsson and Simon (1993) cite verbalization studies which also indicate separate encodings in STM of visual versus oral or symbolic information. In particular,

they cite an investigation by Schuck and Leahy (1966) in which two groups of subjects, verbalizing and non-verbalizing, were presented first with complete images, followed by fragmented versions of the originals and asked to identify the missing fragments. Verbalizing subjects failed to mention significant missing fragments. Subjects who traced the missing fragments on an outline of the image were much less prone to these omissions. Schuck (1973) conducted subsequent experiments in which the verbalizing subjects were given additional time to formulate their descriptions. In these experiments, no statistical differences were reported in the quality of the responses. It is significant, however, as Ericsson and Simon (1993) point out, that the verbalizing subjects required additional time to complete the task comparably.

Klatzky (1980b) summarizes several studies which have provided evidence of visual encodings in long-term memory (LTM). Shepard (1967) showed subjects over 600 color pictures and later administered a recognition test. In the test, subjects were shown pairs of pictures, one previously shown and one new and asked to indicate which picture was one of the originals. The recognition score was 97 percent, more than that scored for similar tests of word recall. Standing, Conezio and Haber (1970) extended the experiment by showing subjects slides of 2,560 pictures for 10 seconds each. On recognition tests administered later, subjects scored 90 percent, again higher than comparable tests administered for words. As Klatzky (1980b) notes, the high recognition levels of pictures relative to words found in both studies suggest a special encoding for visual detail.

The idea that separate codes for visual and verbal information exist in LTM is referred to as the dual-code theory and has been intensely debated in the literature. Klatzky (1980b) cites Paivio (1971) as one of its principal advocates. The theory assumes two systems for storing information in LTM: the verbal system, specialized for words and language, and the imaginal system, specialized for mental images and pictures. The two systems are assumed to be strongly interconnected such that an idea represented as an image can be converted to a verbal label and vice versa. An alternative to the dual-code theory is presented by what is referred to as the unitary theory, which assumes a single propositional encoding for all information in LTM. Klatzky (1980b) cites support for this view in the research of (Pylyshyn 1973; Bower, Karlin, and Dueck 1975; Rafnel and Klatzky 1978). Klatzky (1980b) provides a thorough review and analysis of the two theories. She argues the validity of aspects of each theory and concludes that the debate between the two camps has served to constrain the extremes of either theory. She concurs with others (Kieras 1978; Kosslyn and Schwartz 1977) who suggest that two types of propositions may exist, one which represents visual detail, another for semantic interpretation.

As stated, there is little disagreement in the literature that a special visual encoding exists in STM. The existence of separate visual and verbal codes in LTM is still debated. However, unless one accepts only the most extreme view of the unitary theory, some distinction between the storage of visual and verbal information in human memory seems evident.

Education and Learning: Cognitive Styles

Research in education has shown that individuals tend to exhibit cognitive differences in learning styles and aptitudes, particularly with respect to verbal and non-verbal ability. As cited in Yalow (1980), the non-verbal side has been termed spatial (Butcher 1968; Vernon 1950), figural (Guilford 1967) or fluid-analytic intelligence (Cattell 1963, 1971). Yalow (1980) also refers to research, including (Cronbach 1957; Cronbach and Snow 1977; Snow 1977), which shows that the relationship between learner aptitudes and learner outcomes varies with the type of instruction. This observed relationship is referred to as aptitude-treatment interaction (ATI). Yalow (1980) examined the relationships among aptitudes, instructional supplements, and learning outcomes. Students were first tested on overall, as well as verbal and spatial, abilities. In the experiments, students were given instructional materials on which they were tested. Some students were given supplemental instructional material in the format for which they had tested strongest; others were given no supplemental material. On post-tests given immediately after presentation of the material, students who were given the supplemental instructions scored consistently higher. On tests administered up to a week later, students who were not given the supplemental instructions showed higher retention of the material. According to Yalow (1980), the implications of the study are that when students must work harder, i.e., employ a cognitive style in which they are weak, they retain the material longer.

Yalow's study, and the research on which it is based, is significant because first, it assumes a fundamental difference in the way humans process verbal and visual

information. In addition, it indicates the individual differences in how this processing occurs. Much has been written about these assumptions in the field of human-computer interaction (HCI).

Human-Computer Interaction

HCI research particularly relevant to the discussion of cognitive styles originates from the investigation of multimodal interfaces. Coutaz, Salber, and Balbo (1993) note that in psychological literature, a human sensory channel such as the visual, auditory or tactile, constitutes a modality. They define a multimodal computer interface as one which, in addition to employing multiple sensory channels or "modalities," is distinguished by its ability to provide a higher level of abstraction in the interaction process. They contend that a primary means of providing this additional level of abstraction is through the use of natural language. Shneiderman (1984) also states the advantages of natural language in expressing many of the complexities of human-computer interaction. Numerous studies have investigated the benefits of combining natural language with other modalities, including visually oriented direct manipulation techniques such as touching and pointing (Stock 1994; Cohen et al. 1989; Biermann, Fineman, and Heidlage 1992; Burger and Marshall 1993). Each of the studies lends support to the hypothesis that the two interaction techniques, natural language and direct manipulation, can be used synergistically to compensate for the others' weaknesses and hence produce an interaction technique that is stronger than the sum of its parts.

Stock (1994) argues that natural language can convey complexities that cannot be conveyed through touch or pointing alone. This is not surprising if one considers Shneiderman's guidance that the use of direct manipulation is appropriate when there are a limited number of selectable objects or actions which must be displayed on the screen (Shneiderman 1984). Cohen et al. (1989) argue that natural language is better-suited for describing actions or objects that cannot be referred to directly. Particularly, the use of quantifiers, pronouns, definite noun phrases and tense allow, indeed require the use of context, which makes the interaction more efficient. Finally, a study described by Krause (1993) showed that when users navigating a hypermedia network became completely lost, they almost always reverted to natural language as a means of finding a mode of exit.

While natural language offers the expressive power to convey certain complexities, direct manipulation offers the user a sense of visual immediacy and direct control. Together, natural language and direct manipulation produce the effect of encouraging user navigation and exploration and reducing ambiguity in processing both input and output. For example, a visual picture annotated with natural language text is more explicit to the user, leaving less opportunity for misinterpretation than the picture or text alone. Likewise, touch and pointing can be used to increase the interpretation capability of the natural language component. The Al Fresco system, described by Stock (1994), offers one example. The system is implemented on a videodisc with a touch screen and displays sacred scenes with stereotypical events and objects. It places objects or events in the foreground of the display to indicate to the user what

objects/events are indexed and can be selected for a natural language text description. Touching the image also helps the natural language component to resolve pronoun referents. For example, if the user points at an image and asks, "Who is this?", it is clear that the pronoun refers to a person and exactly which person. Touching an event image and asking "Who takes part in this?" helps to clarify to the system that the referent is indeed an event. The Concierge system, also described by Stock (1994), combines the two modalities to reduce ambiguity in an even more interesting way. It displays to the user a pictorial representation of its interpretation of the user's input and how it fits into the dialogue, allowing the user to select, by touch or pointing, coreferences or alternative representations, thereby directly intervening and correcting misinterpretations if they occur.

To summarize, combining multiple modalities seems to offer the user a more complete and natural interaction with the computer system. The research presented clearly assumes that users employ at least two separate modes of cognitive processing, one language-oriented and one visuospatially-oriented, and that each mode offers particular advantages and disadvantages. In addition, the research demonstrates that restricting the modalities offered by the computer system produces a less natural interaction, which ultimately increases the user's cognitive load.

Cognitive and Perceptual Differences for Individuals with Sight Loss

Research reviewed thus far has focused on the general population, rather than the particular abilities and needs of individuals with sight loss. Clearly, persons with

visual impairments suffer a disadvantage when accessing information which is highly visuospatial in nature. The purpose of the preceding discussion was to help establish the difficulty of such access for sighted individuals as well. Research specific to the cognitive and perceptual abilities of individuals with sight loss should also be presented.

Spatial and Motor Development

The representation of spatial knowledge is perhaps one of the most significant differences between sighted individuals and those with visual impairments. Several studies have been performed on the motor and spatial abilities of infants and toddlers with visual impairments. Ochaita and Huertas (1992) review some of these studies (Bigelow 1986; Ferrell 1986; Fraiberg 1977; Griffin 1981). Fraiberg (1977) and Griffin (1981) found problems in the acquisition of postural control and body rotation in infants with visual impairments. In later stages of growth, children with visual impairments develop locomotor abilities, such as creeping, crawling and walking, much later, an average of 13 months for crawling and 19 months for walking, than sighted infants (Ferrell 1986; Fraiberg 1977; Griffin 1981). Researchers have attributed this delay to the fact that children with visual impairments develop the skill of directed reaching and, therefore, establishing object permanence much later. Because they cannot see the objects in their environment and hence, do not know they exist, they cannot reach for them. This means that the development of the concept of object permanence, necessary to begin crawling and walking, is also delayed in children with visual impairments (Bigelow 1986; Fraiberg 1977).

Cognitive Development

Developmental psychologists have established three cognitive stages through which children pass in developing spatial representations of their environment (Hart and Moore 1973). In the first stage, ages 4-7 years, children use egocentric systems of reference in which elements of an environment are organized only in relation to the themselves, with no other topography imposed. During the second stage, ages 7-11 years, children begin to differentiate clusters and subgroups, but the relationships between subgroups remain primitive. In the third stage, 11 years and above, children develop abstractly coordinated reference systems in which different subgroups and clusters are related with accurate spatial relations, such as Euclidean and projective. In a study of how children with visual impairments progress through these stages, Dodds, Howarth, and Carter (1982) determined that most 11-year old children with sight loss are still in the first, egocentric stage of spatial representation of their environment. In addition, the study showed that adults and adolescents with visual impairments have difficulty in reaching the last stage of development, in which a known space can be represented through a coordinated reference system. However, the study showed that, though not in the majority, some adolescents and adults with visual impairments could represent their surroundings in an organized, coordinated manner. In addition, all study participants improved their abilities with increased familiarity with their surroundings.

Ochaita and Huertas (1993) performed a study, based on Hart and Moore's stages, in which they examined the abilities of people with visual impairments, at each

stage, to learn a variety of geographical routes. They found that at all stages of development, it was not the size of the route which determined how well the individual could learn the route, i.e., develop a spatial model of the route, but the complexity of the route. The study also showed that while increased familiarity with the route helped participants to deal with more complexity, this "learning" was insignificant compared to the individual's cognitive stage of development. Individuals in the latest stage of cognitive development performed significantly better than those in earlier stages, regardless of familiarity with the route. Also, the study verified that most individuals with visual impairments did not progress to this third stage of development until well into or past adolescence, at age 17 or beyond. The researchers drew a conjecture from their results which was pertinent to the considerations of user interface designers, i.e., the development of the ability for abstract and propositional reasoning which occurs at the age of adolescence may contribute to their superior performance in learning the route. In other words, the verbal reasoning which is developed in adolescence may have helped to rectify some of the problems in understanding complex visual spaces due to lack of vision.

Perceptual Abilities

In a discussion of their examination of the auditory functioning of subjects with and without sight loss, Arias et al. (1993) begin by noting informally that most people with visual impairments unconsciously and intuitively produce sounds, such as tongue clicks, snaps, hisses, or verbalizations, as they move to gather spatial information,

whether in a new or familiar environment. Indeed, the authors reviewed many studies which examined the issue of whether subjects with visual impairments exhibit superior nonvisual abilities, particularly auditory, over sighted subjects. Overall, the studies have shown that people with visual impairments possess superior abilities in certain areas. These include auditory localization (determining the location of an audio source) (Rice 1969, 1970), tests of chord analysis (Pitman 1965), memory of melody (Drake 1954), and discrimination of loudness and sound patterns (Stankov and Spilsburg 1978). However, they performed worse on tests to discriminate pitch and rhythm (Juurmaa 1967) and maintaining and judging rhythm (Stankov and Spilsburg 1978).

In the study performed by Arias et al. (1993), the auditory functioning of eight subjects with visual impairments, highly skilled at detecting obstacles, and eight sighted subjects was compared through a series of tests. The tests measured the subjects' abilities to discern pure tones, complex tones, and phonetically balanced probe words (a speech audiometry test). The differences for each test were statistically significant, favoring the subjects with vision loss. The results also showed that the subjects with vision loss may process auditory information through a different pathway in the brain. For further details, see (Arias et al. 1993).

The results of an effort to use tactile pictures, i.e., thermoforms, to communicate the work of visual artists to people with visual impairments should also be considered (Hinton 1991). Many previous efforts to convey the visual arts were based on the assumption that verbal descriptions of a work, accompanied by tactile representations of individual objects within the work, could best convey the overall

meaning. Informal responses to Hinton's thermoforms, which represented entire scenes in works such as paintings or watercolors, suggest that people with visual impairments may understand more about a complex scene through a full tactile representation, although care must be taken at positions in which there are sharp changes in the scene. It should be noted, however, that no formal tests have been reported in conjunction with Hinton's work.

Adventitious Versus Congenital Sight Loss

Finally, it is important to mention research investigating the differences between individuals with congenital versus late onset or adventitious vision loss. Several studies have examined issues pertaining to differences in psychosocial adjustment with respect to the age of onset of vision loss. These include (Beggs 1992; Crudden 1997; Dale 1992; Hudson 1994; and Resnick 1983). Welsh and Tuttle (1997) provide a thorough review of research pertaining to this issue. Though the psychosocial differences are interesting, potential differences in cognitive and perceptual abilities are particularly relevant to this investigation. McLinden (1988) performed a meta-analysis of research examining spatial task performance among individuals with sight loss. The results of the study showed that individuals with early onset blindness performed significantly worse than subjects with late onset vision loss or sighted subjects, indicating that visual experience is correlated with enhanced spatial skills. Welsh and Tuttle (1997) also review research pertaining to the cognitive abilities of individuals with congenital versus adventitious vision loss. They argue that the necessity for a

person with congenital blindness to rely on tactile, verbal, and other auditory cues impedes the cognitive process of integrating parts into a whole. Since this integration process is critical to spatial reasoning, Welsh and Tuttle (1997) contend that many individuals with congenital sight loss may possess deficiencies that cause functional difficulties in areas such as mobility, leisure, or work.

Summary

This research examines the use of a displayless spoken language interface as a means of alternative nonvisual access to spatial data presented through a GUI. Displayless interface technology has been developed and tested primarily for the sighted population. Current GUI access technology was developed and tested for users with visual impairments. Therefore, any effort to employ displayless technology as a component of a solution to the GUI access problem should be based on the fullest possible understanding of the cognitive capabilities and limitations of each population. One issue common to both is the difficulty of using language as the single mode of access to visuospatial information.

To review, neurological and psychological research performed on the general population indicates fundamental differences in human processing of verbal and visuospatial information. Research in education has shown two distinct cognitive modes of learning, as well as interindividual differences in reliance on those modes. HCI research in multimodal interfaces argues that at least two cognitive modes, verbal and visuospatial, can and should be used synergistically in the interface to compensate

for the weaknesses of either alone. This idea is indirectly supported by research concerning the perceptual and cognitive abilities of people with visual impairments. This research suggests certain cognitive limitations, particularly in the representation of spatial knowledge, as well as ways in which individuals with sight loss may compensate for these limitations, e.g., through the use of language and auditory or tactile cues.

GUI Access

This section reviews the current state of GUI access technology. Early obstacles to GUI access are outlined. Current research in GUI access is reviewed.

Early Obstacles to GUI Access

Three early obstacles to the use of GUI's by users with visual impairments have been surmounted in varying degrees (Boyd et al. 1992). The first of these was presented by the use of a pixel buffer rather than a text buffer. Screen-reader software developed for command-line interfaces accessed the information on the screen from a text buffer and transmitted it to a speech synthesizer or some other type of accessible output device. This software could not access the pixel buffer used by a GUI. The problem has been solved by products which can intercept textual information from the pixel buffer before it is displayed on the screen. Berkeley Systems developed an access interface for the Macintosh which was the first commercial application of this approach (Boyd, Boyd, and Vanderheiden 1990). Another impediment was introduced by the use of graphical icons to present information. Interception-based software addresses this problem by recognizing and tracking the location of the icons. The software

captures the text associated with the icon and transmits it to a speech synthesizer which can then read the information to the user (Boyd, Boyd, and Vanderheiden 1990). The third obstacle, the use of the mouse, has been circumvented by substituting the manual mouse functions with keystrokes on the numeric keypad (Boyd et al. 1992).

Open Issues for GUI Access

Although the interception strategy has made GUI's more accessible, it provides only minimal access at best and certainly does not afford users with visual impairments the full benefits offered to sighted users. The major weakness of this approach is that it attempts to apply a solution developed for a speech-based modality, i.e., screen-reading software for command line interfaces, to a modality with an underlying display metaphor which is visually and spatially based. Vanderheiden and Kunz (1990) argued that speech-based access alone cannot provide the navigational capabilities for scanning and browsing, or full access to spatially related information. Also, translating icons to speech lessens their intuitive benefits and contributes to auditory overload.

Boyd, Boyd, and Vanderheiden (1990) defined three stages of accessibility to GUI's, each at an increasingly sophisticated level. The first two stages encompass the minimal level of access offered by the interception strategy. At the third level, Boyd, Boyd, and Vanderheiden (1990) advocated the integration of multiple nonvisual communication channels, including speech output, voice recognition, tactile output, haptics, and auditory cues.

Current GUI Access Research

Current GUI access research seeks to address the limitations posed by the interception strategy and to extend the level of access beyond the first two stages defined by Boyd, Boyd, and Vanderheiden (1990). This includes the use of both non-speech audio cues and tactile output modes.

Non-Speech Auditory Cues

The Mercator project, an interdisciplinary research effort at the Georgia Institute of Technology, has investigated different strategies for providing access to the X Windows environment (Edwards, Mynatt, and Rodriguez 1993; Mynatt and Weber 1994; Edwards, Mynatt, and Stockton 1994). This project offers a unique approach to the GUI access problem for several reasons. First it addresses the navigation problem by discarding the spatial representation of the GUI in favor of a tree structure, based on the X-widget hierarchy. Users simply traverse the tree to navigate the interface (Mynatt and Weber 1994). This representation is quite natural because it exploits the predefined parent-child relationships between objects in the visual display. It also precludes the necessity of dealing with problems such as occluded or iconified windows, which are simply artifacts of the visual display.

In addition to translating text into speech, Mercator employs non-speech audio cues similar to the auditory icons in Gaver's SonicFinder (1989) to convey symbolic information. Gaver (1989) defined non-speech auditory icons to represent interface actions in the Macintosh File Finder. These auditory icons are naturally occurring,

everyday sounds. For example, in Mercator, the sound of a chain-pull light switch represents a toggle button, while a muffling filter applied to this auditory icon indicates that the button cannot currently be selected. In addition, a more abstract representation akin to the "earcons" defined by Blattner, Sumikawa, and Greenberg (1989), is employed to augment navigating the hierarchical tree. Earcons are constructed from differing rhythmical combinations of musical timbres. For example, the octaves of a piano are used to indicate the user's relative position within a menu or list (Edwards, Mynatt, and Stockton 1995).

To review, Mercator offers the advantage of relying less on an exact translation of the screen contents, yet providing the user with certain benefits of a direct manipulation interface, including the intuitiveness of icons through audio cues. This means that it requires less visualization of objects and concepts which are not inherently spatial, but are represented spatially in the interface, such as a file menu or dialog box. It does not however, provide any additional means of representing information which is inherently spatial, such as geographical or symbolic maps. Also, Boyd, Boyd, and Vanderheiden (1990) argued that the use of spatial concepts need not be abandoned entirely since they can convey semantic information such as relative importance, similarity, and group membership, but their use may require modification for people with visual impairments.

Tactile Interaction

Tactile devices offer another, perhaps more natural, means of representing spatial information. Burger (1994) notes that refreshable Braille displays, though expensive, provide an acceptable tactile method for presenting textual information, but are not well-suited for conveying graphics. He offers the same criticism for a commercial device called the Optacon which uses technology similar to the refreshable Braille device to produce tactile images of English letters and words. This device has been used as a computer terminal for the InTOUCH Macintosh access product (Berliss 1993), but is no longer widely used.

Hill and Grieb (1988) developed a touch device for presenting spatial concepts in a nonvisual interface. Their study explored a multimodal approach to screen representation, using a workstation with a touch-sensitive pad and a speech synthesizer. Subjects used a stylus to touch and manipulate data presented on the surface of the pad. Given two tasks to perform, locating an area on a page as well as a standard editing task, the subjects performed almost 50% better using the spatial device. Similar research was conducted by Burger et al. (1993), who developed a learning tool for children with visual impairments which associated tactile images and sounds. During testing, the children were able to examine tactile images represented on the touch device and to directly locate some portions of the image. Finally, some commercial products are available, such as Nomad (Uslan, Schreier, and Meyers 1990) which allow presenting tactual representations of geographical maps.

Clearly, if a multimodal approach, employing a tactile output modality, is to be applied to the GUI access problem, a hands-free input mode will be desirable. As mentioned, Boyd, Boyd, and Vanderheiden (1990) advocated the use of voice recognition for achieving the third and most sophisticated stage of access. This means that research in spoken language technology, particularly displayless interfaces, should be examined.

Displayless Interfaces

Recent improvements in digital device technology have contributed to advances in both speech synthesis and recognition. Speech processing hardware, based on digital signal processors (DSP's), is now available for personal computers and workstations, simplifying the development of application programs. These advances, along with improvements in the mathematical modeling of the speech sound, have made synthesized audio as well as speaker-independent, continuous speech recognition on restricted vocabularies possible at the desktop level. Many human factors issues relating to the use of speech-based interfaces remain unresolved, however. Bradford (1995) gives a thorough overview of these issues, the most pertinent of which for this discussion include appropriateness of application for speech, verbal cognition, auditory overload and prosodics.

Applications Appropriate for Speech Interfaces

Speech-based interfaces are particularly appropriate for certain situations. Shneiderman (1992) details four situations for which speech is appropriate. First, tasks

in which the hands are busy are often better performed with the use of voice input. Shneiderman (1992) gives the example of an inspection worker on an assembly line. An individual with vision loss using a tactile computer output device provides another example. Second, situations in which the eyes are busy or vision is restricted (through the environment or an impairment) make speech input and output more convenient and in some cases, essential. Third, speech is more desirable in situations that require mobility. Finally, when computers are used in harsh environments in which keyboards and screens cannot be used, a speech interface may be required.

Many applications which meet Shneiderman's criteria are found in military environments. Army applications of displayless technology include Command and Control on the Move (C2OTM) and the Soldier's Computer (Weinstein 1994). C2OTM is an Army program designed to ensure the mobility of command and control for future needs. For mobile users whose hands and eyes are often busy with other urgent tasks, typing is not an optimum mode of input. Voice input would provide a more convenient means of transmitting reports or accessing battlefield situation information. The Soldier's Computer is an Army Communications and Electronics Command (CECOM) program designed to address the needs of the modern soldier (Weinstein 1994). Voice input and speech output are critical for this application since transporting and using a keyboard and terminal would be awkward for the foot soldier. Air Force applications for speech-based interfaces include the use of speech output for critical warning messages as well as voice control of radio frequency settings in fighter cockpit applications. Weinstein (1994) notes that the Federal Bureau of Investigation

(FBI) has identified applications, comparable to those of the military, requiring speech access. Similar to the Soldier's Computer, the Agent's Computer would be a portable device with particular functions of interest to the agents, including data or report entry, covert communication, and rapid access to map and direction information.

Another example, surprisingly similar to the military-based applications, can be seen in research conducted by Loomis et al. (1994) which investigated the use of a Geographic Positioning System (GPS) and Geographical Information System (GIS) as a navigational aid for travelers with visual impairments in unfamiliar environments. To investigate the concept, Loomis et al. developed a prototype navigational tool, consisting of three modules, a GPS, a GIS, and a user interface. The GPS module uses a hand-held satellite receiver to determine the traveler's relative position and orientation. The GIS module contains a spatial database of the environment, linking spatial information about objects, such as shape or location to nonspatial properties such as surface traffic. The GIS can provide spatial layout and route planning data or compute information such as the number of objects of a given type on a route. All of this information is conveyed to the traveler via the user interface. At the time of publication, only the output interface was developed. It offered a choice of binaural headphones with speech and non-speech audio cues (similar to those used in Mercator) or a pure speech-based auditory display. The authors planned to implement either keyboard or speech input. Clearly, speech would offer the more convenient input mode. The application meets many of Shneiderman's criteria for use of a speech interface, addressing the needs of mobile users whose vision is restricted, and who are

operating in circumstances in which a keyboard and screen are not appropriate. Also, if a tactile map were added to the output interface, busy hands would make voice input essential.

Many of the applications described involve the necessity of accessing spatial and geographical data or data which may be presented spatially through a graphical interface. Bradford (1995) includes such applications in a list of those for which speech is not the optimum medium. Nonetheless, for many of these applications, e.g., GUI users with visual impairments or mobile users accessing GIS maps in harsh environments, speech becomes, by necessity, the most desirable of the available choices. Therefore, it is important to determine how to best employ speech in these situations. This leads to a discussion of the importance of developing a better understanding of verbal cognition and how it relates to visual cognition.

Verbal Cognition

Bradford (1995) argues that verbally based cognition as opposed to visually based cognition must be examined from a human factors perspective if speech-based interfaces are to enjoy widespread use. He argues that some users may be more natural verbal and acoustic thinkers, while others may function better at visual and spatial thinking. Research conducted by (Yalow 1980) supports his argument. Robertson (1985) reviews other research regarding the existence of differing cognitive styles and strategies for information processing.

In addition to differences in users' styles of cognition, some problem domains lend themselves more to a verbal versus visual style and vice versa. Nonetheless, in certain situations, access to visually based information through speech may still be desirable or even necessary. An ongoing investigation at the Naval Research Laboratory is examining issues in using language to access spatial data in a GIS (Marsh, Wauchope, and Gurney 1994). The project has entailed development of voice input and speech output as part of a multimodal interface to the GIS. The GIS database contains a complex set of spatial relationships describing a large geographical area in Germany and provides a rich set of data for the investigation. Thus far, results have indicated the benefits of a multimodal interface over the use of graphics or language alone.

Auditory Overload

Although it can be intensified by the user's cognitive style as well as the nature of the application, auditory overload presents a challenge for any interface which uses only speech, and no visual display, as the mode of output. Sufficient information must be presented so that the user can make choices to perform tasks; however, too much information can easily overload the user's short-term memory and cause frustration. Finding the proper balance between these conflicting goals is difficult. In developing a displayless interface to an on-line air travel information system, Zue et al. (1994) employed an extensive interactive dialogue with users to limit the scope of information requested in order to avoid aural information overload. The effort necessary to

conduct such a dialogue, however, could also contribute to user frustration. In addition, it would likely render the solution inadequate in time-critical situations.

Prosodics

As stated, the proposed research will examine the effects on the nonverbal quality of speech produced in accessing spatial data through spoken language. Much information is contained in spoken language beyond simply the words or word sequences which can be detected by the recognition component or spoken by the synthesizer. This information is referred to as the prosodics of spoken language. As an example, pauses, intonation and register in the computer speech output convey meaning to the user. Similarly, prosodic information contained in the user's speech, such as the change in duration of phonemes or the presence of embedded silences, can also convey meaning. Moore gives the example of the sentence, "What do the fare codes BH and K mean?" versus "What do the fare codes B,H, and K mean?" (Moore 1994, 269). The two sentences differ prosodically: when spoken, the second sentence would contain an embedded pause between the letters 'B' and 'H'. The sentences also have two entirely different meanings. Prosodic information has been used to reduce syntactic ambiguity in sentence parsing (Bear and Price 1990; Price et al. 1991) as well as to detect sentence phrase boundaries (Wightman and Ostendorf 1994).

In addition to pauses, prosodics can include sentence and phrasal prominence or stress. This refers to a speaker's emphasis of certain words within a sentence, manifested in acoustic factors such as vowel duration and word intensity. This

emphasis often conveys semantic significance. For example, consider the following sentences (words of emphasis are italicized): "Call in the *Monday* report." versus "Call in the Monday *report*." The first sentence indicates that the day of the report is significant. The second sentence indicates the report itself is the item of significance. Algorithms have been developed which detect this type of information (Chen and Withgott 1992; Campbell 1992).

Prosodics require further study for an additional reason. Research performed in both laboratory and everyday settings has demonstrated that human speech changes in situations of stress and emotional tension (Scherer 1981). In such situations, the emotion which computer users are most likely to experience is frustration. Current interfaces can get caught in a cycle of misrecognition due to user frustration, followed by increased user frustration, followed by misrecognition, resulting in a spiraling degradation of system performance. As Bradford (1995) points out, a better understanding of the acoustical correlates of prosodics, register, and emotion could be beneficial in resolving these misrecognition errors.

Summary

This section has reviewed general issues pertinent to the development and use of displayless interfaces. It seems clear that speech interaction will be needed for users with visual impairments to advance to the next level of access to GUI's defined by Boyd, Boyd, and Vanderheiden (1990), yet these interfaces introduce unique problems. In particular, how can nonvisual, spoken language access to spatial or graphical data be

optimally provided? This problem is also relevant for sighted users of this technology, particularly those who must employ the technology in restrictive environments to access geographical or other spatial data. To improve the quality of interaction for both categories of users, this issue should be investigated. This research examines the effects of using a displayless interface to access visuospatial information. As indicated in neurological, psychological, educational, and HCI literature, this task will tend to increase the user's cognitive load and thus, difficulty in the interaction. Hence, the research examines specifically the effects of such access on the prosodics of the speaker's utterances. The following section discusses the area of speaker prosodics in greater depth.

Speaker Prosodics

Speech prosody has been studied by practitioners in numerous disciplines, including linguistics, psychology, psychiatry, and digital speech recognition and production. Study in each discipline has been motivated by different goals, e.g., linguists are interested in the study and teaching of languages, whereas researchers in digital speech production are interested in reproducing human prosody in digital speech. Certain concepts and definitions, however, are common to all disciplines. These definitions, along with certain language-independent features of prosody, are presented in this section, followed by a review of relevant prosodic research specifically in the area of automatic speech recognition. The section concludes with a review of

research on the effects of psychological arousal on human speech production, followed by presentation of the research hypothesis.

Wightman and Ostendorf (1994) define prosody as pertaining to attributes of the speech signal beyond the spoken words, such as timing and fundamental frequency (F0) patterns. They also note that prosody is often termed suprasegmental information because it represents more than what can be found in a single phone-sized segment. Cruttenden (1986) details suprasegmental features which constitute prosody to include pitch, loudness, stress, accent, pauses, intonation, and rhythm. Some of these features, such as accent and intonation, vary according to language. Certain language-independent features, however, can be identified.

Vaissiere (1983) categorized various language-independent prosodic features and reviewed linguistic research regarding these features. The features, summarized below, include pauses, F0 features, and durational and intensity features.

Language-Independent Prosodic Features

Pauses

Speakers tend to pause to breathe at the end of large units of information such as clauses or sentences. In addition, pauses between sentences tend to be longer than pauses within a sentence (Goldman-Eisler 1972). Another type of pause, unrelated to grammar, referred to as the hesitation pause, is likely to occur in spontaneous speech. The duration and frequency of this type of pause depends on several variables, such as speech rate and speech mode. At faster speaking rates, hesitation pauses tend to be

suppressed (Grosjean and Collins 1979); they occur around grammatical junctures less frequently in spontaneous speech (Goldman-Eisler 1968) and more frequently in read speeches (Duez 1985), and vary depending on the emotional state of the speaker (Fairbanks and Hoaglin 1941). Either type of pause can be unfilled, i.e., silent, or filled, in which some type of non-verbal voicing occurs such as "um"or "er".

F0 Features

Several properties of F0 contours have been observed in short, simple utterances spoken without a pause. These include the declination tendency, normal frequency range and control, and rising versus falling F0 movements.

Declination Tendency

This refers to the overall tendency of the F0 curve to decline over time, even though local rises and falls may occur. Some physiological explanations have been proposed and argued in the literature (Liberman 1967; Hixon, Klatt, and Mead 1971). Regardless of explanations, the declination is easily observed, but the rate may vary.

Normal Frequency Range and Control

The range of F0 variations tends to lessen over time. In other words, the local F0 minima and maxima decrease over the length of a simple utterance. Similar to the declination tendency, this tendency has been attributed, to some extent, to physiological factors (Ohala and Ewan 1973). The largest rise in F0 typically occurs in one of the first 3-4 syllables of a sentence. Also, the lowest F0 in the sentence usually occurs

when the speaker ceases voicing. However, speakers may suppress this tendency, for example, to delineate a declarative sentence from a yes/no question (Thorsen 1980).

Rising versus Falling F0 Movements

Rises and falls in F0 can signify information about phrase or sentence endings or sentence types, as mentioned. F0 contours used in this way, such as in delineating yes/no questions from declarative sentences or marking phrase and sentence endings, are referred to as boundary tones. Beckman and Pierrehumbert (1986) define two rising and falling F0 boundary tones (H and L) which mark intermediate phrases and two (H% and L%) which mark major intonational phrases. They give the example of the phrase, "a round-windowed, sun-illuminated room" (Beckman and Pierrehumbert 1986, 268). The phrases 'round-windowed' and 'sun-illuminated' are each intermediate phrases with a rising (H) F0 on 'round' and 'sun' and a falling (L) boundary tone on 'windowed' and illuminated'. The entire phrase constitutes an intonational phrase with an L boundary tone on 'room'.

F0 contours can also be applied to words to convey relative sentential stress or prominence. For example, in the sentence, "The report is here", a rising contour on "report" indicates prominence on this word, suggesting the report itself is important, while a rising contour on "here" would place stress on this word, indicating importance of the location of the report. These intonational contours, known as pitch accents, are defined in (Bolinger 1958). Pierrehumbert and Hirschberg (1990) define six pitch

accents which are combinations of high and low pitch targets, consisting of a main tone and an optional leading or trailing tone.

Durational Features

Final Lengthening

This feature refers to the tendency of speakers to lengthen final components of an utterance, particularly the last vowel occurring before a pause. Lengthening without a pause is also used to indicate the end of a word or phrase (Vassiere 1983). Final lengthening is also referred to as preboundary lengthening, particularly in reference to the detection of prosodic phrase boundaries in the automatic recognition literature.

Other Durational Lengthening

Lengthening of a non-final syllable is often used to show emphasis or contrastive stress (Vassiere 1983). Also, speakers may slow their speaking rate to stress a word, sentence, or clause (Vassiere 1983). These types of non-final lengthening are also referred to in the automatic recognition literature as durational lengthening.

Prosodic Pattern Detection Algorithms

Algorithms to detect prosodic patterns in speech have addressed several problems, including phrase structure recognition, tune recognition, and prominence or stress detection. Phrase structure recognition algorithms are necessary because it is often difficult to determine the end of one utterance and the beginning of the next in

spoken language recognition. Prosodic phrase boundaries can serve as important cues. Also, prosodic phrase boundaries have been used to disambiguate syntactically ambiguous sentences (Price et al. 1991; Wightman, Veilleux, and Ostendorf 1991). Tune recognition algorithms address problems such as determining boundary tones to discern yes/no questions. Stress or prominence detection algorithms address the problem of detecting the relative prominence of a syllable or word within a sentence. Work in each of these areas has tended to concentrate on the use of different and limited prosodic cues. The algorithms are reviewed briefly.

Prosodic Phrase Detection

Much of the work in prosodic phrase detection has relied on the use of F0 contour analysis. Huber (1989) developed an algorithm based on the assumption of overall F0 contour declination. A shift upward is assumed to start a new declination line, and hence a new phrase boundary. Shimodaira and Kimura (1992) used dynamic programming to find the optimum phrase segmentation from a set of F0 phrase templates derived through clustering. The approach was used for speaker-dependent cases only. Nakai, Shimodaira and Sagayma (1994) extended the work of Shimodaira and Kimura (1992) to speaker-independent continuous speech. Using eight speaker templates, their approach successfully detected approximately 83% of prosodic phrase boundaries; however, their reported insertion error (false detection rate) was quite high, almost 50%. Okawa et al. (1993) used vector quantization of F0 patterns as well as phonemic characteristics to automatically detect phrases. In their experiments, they

used two approaches, one using no grammar and a second in which they used a simple bigram model grammar. Using no grammar, their algorithm achieved 72.8% accuracy in detecting phrase boundaries with an 11.4% insertion error. Using the simple grammar, 78.7% accuracy in phrase boundary detection was achieved with no insertion error.

As stated, all of the previous studies relied primarily on F0 contour analysis for phrase detection. Rather than assume the significance of one feature, Wang and Hirschberg (1991) applied Classification and Regression Tree (CART) techniques (Brieman et al. 1984) to the DARPA Air Travel Information Services (ATIS) database to determine the predictive power of various features in identifying phrase boundaries. Prior to the experiments, the speech corpora was labeled prosodically by hand, marking the type and location of phrase boundaries and presence of pitch accents, for comparison and analysis. Features examined in the experiments included part-of-speech, as well as intonational aspects of the utterance, i.e., pitch accents, boundary tones, utterance and phrase duration, and prior boundary location. Four sets of experiments, using differing combinations of features, were performed. In the experiments using only boundary location and pitch accents, agreement of phrase boundaries with the hand-labeled data was almost 90%. The lowest rate of agreement quoted for any of the experiments was 88%. Wang and Hirschberg (1991) concluded that considerable redundancy exists among the features useful in phrase boundary prediction. They give no information on false detections, however.

Tune Recognition

Automatic tune recognition algorithms handle classification of boundary tones as well as pitch accents. Using rule-based techniques, algorithms which detect yes/no questions from F0 contours have yielded high accuracy (Waibel 1988; Daly and Zue 1990). These algorithms, however, are strictly limited to this problem. Hidden Markov Model (HMM) algorithms offer an approach which could be used for both classifying boundary tones or pitch accents. HMM's have been employed for related problems such as contour classification (Ljolje and Fallside 1987; Butzberger et al. 1990). None of these algorithms, however, use durational cues and are thus limited to intonation pattern classification, rather than prominence or phrase detection.

Prominence Detection

Stress or emphasis detection algorithms detect the relative prominence of a syllable. As Wightman and Ostendorf (1994) point out, in both the linguistics and automatic recognition literature, "stress" is used ambiguously to mean both relative strength of syllables denoted by lexical stress and phrasal prominence denoted by pitch accents. Recent recognition systems model lexical stress directly using separate models for stressed and unstressed vowels. Phrasal level prominence algorithms employ several techniques, including HMM detection of emphasis using frame-based energy and duration features (Chen and Withgott 1992), linear discriminant functions based on syllable-level features (Hieronymous, McKelvie, and McInnes 1992) or frame-level

features (Campbell 1992). Only Campbell (1992) gives detection rates, 72-92% correct with 4-7.5% false detection rate.

Wightman and Ostendorf (1994) seek to address what they view as the shortcomings of many of the above algorithms in using only limited acoustic cues (with the exception of Wang and Hirschberg (1991)) as well as in working with the speech signal directly. They describe an automatic prosodic labeling algorithm which uses multiple acoustic cues and works with the output of a speech recognizer rather than the actual speech signal. Using the phonetic segmentation produced by the word recognizer allows an analysis of phrase-final lengthening and other durational cues. Their algorithm handles detection of both prosodic phrase boundaries and phrasal prominence. Unlike many of the other algorithms described, numerous prosodic cues are used to detect phrase boundaries. These include preboundary lengthening, pauses, breaths, boundary tones, and speaking rate changes. The cues used to detect phrasal prominence include duration lengthening and pitch accents.

The algorithm approaches prosodic labeling as a standard pattern recognition problem, requiring feature extraction and classification. Features are extracted at the syllable or word level and then classified, using a decision tree (CART), to syllable or word level prosodic labels. Different features are extracted for phrase boundary detection versus phrasal prominence detection, as noted above. Once the feature vectors are extracted, the classification module first employs decision trees to determine the relative importance of the cues in the feature vectors for prosodic labeling. A Markov sequence model is then used to determine the most likely labeling

sequence. This takes advantage of certain sequencing likelihoods, e.g., two sentence-level phrase boundaries are unlikely to occur in direct sequence since this would represent a one-word sentence.

To label prosodic phrase boundaries, Wightman and Ostendorf (1994) use the phrase boundary or "break" labeling system defined by Price et al. (1991). This system specifies seven levels of perceived phrase boundaries or breaks, assigned by human labelers. The 0 level is assigned to two consecutive words in which a phonetic reduction has occurred, such as a deleted /h/ in "did he?". A break index of 1 is assigned to a normal word boundary; 2 is assigned to a grouping of words having only one prominence, 3 to an intermediate phrase boundary, 4 to an intonational phrase boundary, 5 to a boundary signifying a group of intonational phrases, and finally sentence boundaries are assigned the label 6. Break indices 4-6 are considered major prosodic breaks, corresponding to what Beckman and Pierrehumbert (1986) define as intonational phrases, while the break index 3 corresponds with their definition of an intermediate phrase.

In detecting prominences, to avoid confusion of pitch accents, typically used to mark prominences, with the boundary tone intonation marker, Wightman and Ostendorf (1994) chose to also model boundary tones, useful for prosodic phrase recognition. They arrived at the following set of intonation markers, "P" for prominent syllables, "s" for unmarked syllables, "BT" for a syllable marked with an intonational phrase (break levels 4-6) boundary tone, and "P-BT" for syllables marked with both a prominence and boundary tone.

The algorithm was tested on two corpora of professionally read speech, which were selected on the basis of the availability of hand-labeled prosodic markers. The ambiguous sentence corpus, developed by Price et al. (1991), contains 35 pairs of syntactically ambiguous sentences read by four professional FM radio announcers. Experiments on this corpus were speaker-independent with three speakers used for training and one for testing. Experiments on the second corpus, a collection of radio news stories read and recorded by one female FM radio, were speaker-dependent. Both corpora were hand-labeled with the 7-level break index and the binary prominence labels.

On the speaker-independent break index labeling experiments, overall accuracy of the algorithm was 55% exact identification and 88% identification within +-1. Assuming break levels 4-6 to be major phrase breaks, their correct detection was 64% with a false detection rate of 6%. For the speaker-dependent break index detection experiments, the average exact identification accuracy was 67% with 89% correct identification within +-1. Major phrases were correctly detected with an accuracy of 78% and a 7% false detection rate. In the speaker-independent prominence and boundary tone labeling experiments, prominences were detected correctly at a rate of 83% with a false detection rate of 14%. On boundary tones, however, the algorithm yielded an accuracy rate of 77% with only 3% false detection rate.

Analysis

Comparing the various prosodic pattern detection algorithms on the basis of accuracy is difficult since detection rates are not given for all, and false detection rates are not given on others. However, of those quoted, the 83% accuracy rate reported by Shimodaira, Kimura , and Sagayma (1992) is somewhat offset by the relatively large false detection rate, 50%. Okawa et al. (1993), using only pitch pattern and phonemic information, quote rates for prosodic phrase detection, 72% correct with an 11.4 % false detection rate (using no grammar), comparable to those of Wightman and Ostendorf (1994), 78% correct with 7% false detection (speaker-dependent corpora), who use multiple acoustic cues. Examining the importance of different cues for phrase detection, Wang and Hirschberg (1990) surmised that many of the cues were redundant. They give no information, however, on false detections.

On prominence detection, Wightman and Ostendorf (1994) did not achieve significantly better results than Campbell (1992), but they claim an advantage of their algorithm, other than performance, is its generality for prosodic pattern detection. Because they use a standard pattern recognition approach with feature extraction and classification, their algorithm can be trained to detect different prosodic features. Clearly, many of the other algorithms, such as those which perform tune recognition, are restricted to certain classes of problems. One significant, but unresolved issue emerging from all the studies is the usefulness of multiple cues in detecting prosodic patterns. In particular, how significantly does the use of multiple acoustic cues improve prosodic phrase recognition or prominence detection?

One final issue should be considered in evaluating the results: some of the studies examined spontaneous speech, e.g., Wang and Hirschberg (1991), although not all, e.g., the ambiguous corpus used by Wightman and Ostendorf (1994). None, however, specifically examined spontaneous speech produced in harsh environments or under conditions of cognitive or other psychological stress, such as the conditions described for the proposed research. The following section reviews research examining the effects of psychological and physiological arousal on speech production.

Acoustical Correlates of Emotion

Interest in the effects of psychological arousal on human speech dates back at least to Darwin (1872). More recent research studying these effects falls into three categories, studies of simulated emotions, studies involving laboratory-induced stress, and studies of real-life situations of danger and stress. Before discussing these studies, some description of observed physiological effects of stress, as described by Scherer (1981), is necessary.

Physiological Effects

Stress-producing situations can affect either the sympathetic nervous system (SNS) or the parasympathetic nervous system (PNS) (Scherer 1981). Increased SNS activity is typically caused by the emotions of fear and anger and has been observed to produce several physiological conditions. Heart rate and blood pressure increase. The rate, depth, and pattern of breathing are altered. A decrease in salivation often occurs as well as sweating of the palms and slight muscle tremor. Conversely, increased PNS

activity causes a decrease in heart rate and blood pressure, a diversion of blood to the digestive tract, and increased salivation. This response is usually the result of a calm, relaxed mental state, but can also occur as a result of feelings of defeat, depression, and grief.

As noted by Scherer (1981), these physiological conditions could potentially produce numerous effects on the quality of speech production. In particular, one might expect the change in respiration and muscle tone to affect the vibration of the vocal cords, the rate of speech, the range of frequencies of vocal cord vibration, and the contour of F0 versus time, among other effects. The following studies confirm many of these expectations.

Studies of Simulated Emotion

Several studies have been conducted in which subjects were asked to simulate certain emotions. The effects of the simulated emotions on speech production were then observed. Two of the earliest and most influential studies were conducted by (Fairbanks and Pronovost 1939; Fairbanks and Hoaglin 1941). For the experiments reported in (Fairbanks and Pronovost 1939), six amateur male actors were asked to read a passage five times, simulating five different emotional states for each reading. The readings were recorded and measurements of F0 were taken from the voice samples, considering all six voices as a single group. In these measurements, fear was associated with the highest median F0, followed by anger, then grief, then indifference. Fear, contempt, and anger produced the greatest mean rate of change in F0. Also, F0

curves showed wider, more rapid inflections for anger and more irregularity of F0 changes for fear.

The Fairbanks and Hoaglin study (1941) examined the durational features of the five simulated emotions from the previous study. This study showed that grief and contempt produced a slower speaking rate. In the case of grief, this was attributed to the prolongation of pauses, especially between phrases. A relatively rapid speaking rate, however, was observed for fear, anger, and indifference.

Williams and Stevens (1972) also studied the vocal quality of subjects simulating certain emotions. To increase the authenticity of the emotions, they enlisted actors trained by the Actor's Studio in "method" acting, a technique which teaches actors to become deeply involved in experiencing a character's actual emotions. A play, written specifically for the experiments, included three characters and several scenarios, each involving expression of a certain emotion with smooth transitions between scenarios. The actors' voices were recorded as they performed the play three times, changing roles for each performance. Detailed acoustic analyses were performed on certain "control clusters," phrases and sentences written to carry a high emotional content. In addition, longer speech samples, including sentences surrounding the control clusters, were taken for analysis.

Mean speaking rates were calculated from the speech samples at points where the emotional content was clearly defined. The speaking rates varied, from highest to lowest, for all three speakers in the same order: neutral, anger, fear, and sorrow. The

speaking rate for sorrow was less than half for the other emotions. This finding agreed with the results of Fairbanks and Hoaglin (1941).

F0 contours of the control clusters were taken by tracing harmonics of narrow-band spectrograms of the utterances. Contours reported for the second speaker showed overall declination of the F0 maxima and minima throughout each utterance, but the shape of the contour differed for each emotion. Contours of neutral control clusters were smooth and continuous with relatively slow changes in F0. Contours from angry utterances were higher throughout and showed greater range of F0, indicating, as Williams and Stevens (1972) suggest, greater use of respiratory and laryngeal muscles. Relatively flat contours were noted for sorrowful utterances with F0 peaks lower than those of neutral utterances.

In addition to the control cluster analyses, measurements of F0 were taken from narrow-band spectrograms of the longer speech samples every 0.15 sec. Distribution curves were created from these measurements to assess the median F0 and to measure the range of F0. A clear distinction was evident among the data for situations labeled "neutral," "anger," and "sorrow". The median F0 was less for sorrow than for neutral and higher for anger than for neutral, with few exceptions. The range of F0 was generally less for sorrow than for neutral and highest for anger. The range of F0 was also, in most cases, greater for fear than for neutral and comparable to that of anger. However, these differences were not consistent for all samples and for all voices. In general, though, the data for the three voices suggested certain trends in F0 for different emotions, as found in Fairbanks and Pronovost (1939). Williams and Stevens

(1972) note, however, that median F0 measurements and ranges alone cannot identify an emotion and should probably be considered together with F0 contours. They also point out the limitations of any study using simulated emotions. Regardless of the level of authenticity sought, certain physiological effects produced by actual emotions may not be reproduced in a simulation.

Studies of Laboratory-Induced Stress

While experiments using simulated emotions present the problem of authenticity of reaction, experiments using laboratory-induced stress present a different problem. In particular, it is difficult to identify a priori the conditions each individual perceives as stressful as well as the level of stress perceived. Therefore, strong individual variations are observed in these studies. Nevertheless, they offer an additional source of information and should be considered. A selective survey of these studies is presented.

Bonner (1943) performed a study using experiments which attempted to induce stress. In one experiment, students in a speech class were required to make an impromptu appearance on a radio broadcast. They were forced to sit in the radio station for some time before the broadcast to absorb the tensions in the atmosphere, and thus heighten the perception of stress. Students were then given a sentence to speak on air, which was recorded for the experiment. For the first three classes, recitations varied, but for the last two classes all subjects spoke the same sentence. For the control-recordings, produced up to one week after the experimental recordings, students were placed in relaxing and comfortable environments and asked to repeat the

sentence spoken on the experimental record. In another experiment, students were blindfolded and subjected to various unpleasant stimuli, such as wet spaghetti or a piece of ice drawn over the hand or placing the hand in a goldfish bowl. Afterwards, subjects were asked to say "Ah" and hold the sound for at least two seconds.

The results of the recordings were analyzed along several dimensions, melody or frequency-rate, rhythm, including hyphae-time (time to speak a syllable) and pause-time (time between hyphae), accent, vowel (measured in terms of amplitude, complexity, and frequency-rate), and consonant (measured by mode of attack and relea0se of syllables). The results, detailed in Bonner (1943), showed wide individual variations in the subjects' responses. Hence, no clear-cut trends were observed in the data. However, the results demonstrated that under the laboratory-induced tension versus the control situation, more individuals demonstrated higher frequency-rates, longer hyphae-time and pause-time, and more who attacked and released the hyphae in a hard rather than easy manner. In addition, Bonner (1943) concluded, as did Williams and Stevens (1972) later, that no single attribute of the speech, nor any single component, such as melody, accent, or rhythm, could be used to determine emotional content of speech, but rather a combination of attributes was necessary.

Studies in which stress was induced by achievement tasks have particular relevance to the proposed research. Scherer (1981) notes that few experiments using cognitive or other achievement tasks have been reported. He cites research by Brenner, Branscomb and Schwartz (1979) in which subjects were presented with standard arithmetic tasks. Using the Psychological Stress Evaluator (PSE), a device

which measures the absence of microtremor as a stress indicator (Holden 1975), they found different PSE measures for difficult and easy parts of the task. Scherer (1981) criticizes their claim of a strong linear relationship between the two, noting that the reported significance level was based on comparing tasks which required no arithmetic processing, such as adding 0 to the end of a number, to any task which required arithmetic operations, making no distinction between gradations of difficulty. Scherer (1981) states that he does not doubt, however, that a relationship exists between the PSE scores and some aspects of the subjects' speech, but that the exact parameters cannot be determined due to the methodological weaknesses of the study.

Goldman-Eisler (1968) studied the relationship between vocal quality and cognitive processing. She examined hesitation phenomena in subjects given tasks to describe or explain magazine cartoons. Three findings of the study were significant. First, pauses occurred less at grammatical junctures in the spontaneous speech produced for the explanations than in speech read from a prepared text analyzed in other studies. Second, longer hesitations usually occurred when the ensuing speech contained increased amounts of information, and third, unfilled pauses occurred more often around creative output than filled pauses. Goldman-Eisler (1968) concluded that pausing might be an attribute in spontaneous speech associated with verbal planning. Results conflicting with those of Goldman-Eisler have been reported (Boomer 1968). Debate of the issues regarding the hesitation phenomena is reviewed by Rochester (1973).

Real-Life Studies

Several studies involving real-life situations of danger and stress have been reported. These studies offer a unique source of information, but, as with the other studies, present certain problems. Chiefly, real-life situations lack the methodological control necessary to duplicate the experiments. Nonetheless, the authenticity of emotion recorded in these studies cannot be captured by any other means.

The most realistic situations studied originate from air-to-ground communications in aviation and space-flight under dangerous conditions. Williams and Stevens (1969) analyzed a recording of a conversation between a pilot and a control tower operator during a flight which encountered difficulties and ended with the pilot losing control and crashing. Narrow-band spectrograms taken from the recording show that the F0 contours increased, becoming irregular and discontinuous, as the fear of the pilot increased. At the end of the recording, when the pilot was in a state of terror, the F0 became quite high, fluctuating widely.

Williams and Stevens (1969) also analyzed a recording of the radio announcer who witnessed the destruction of the Hindenburg. As observed in narrow-band spectrograms taken from the recording, a high level of inflection was present in the voice, indicated by smooth vertical movements of F0. After the crash, the average F0 was much higher with far less fluctuation in frequency. Irregularities present in the contour may reflect irregular breathing or loss of muscle control. The irregularities can also be attributed, according to Williams and Stevens (1969), to the combination of

emotions elicited by the situation, including grief, creating a flatter F0 contour and fear, creating a higher F0.

Williams and Stevens (1981) also cite research conducted by Kuroda et al. (1976) in which a method was constructed to ascertain the emotional state of pilots during aircraft accidents. From spectrograms of pilots involved in crashes, they calculated a vibration space shift rate (VSSR), which signified the rate of change in average F0 under the condition of emotional stress relative to a normal condition. The profiles of these VSSR's were developed for use in determining the relative variation in the pilot's emotional state. Pilots who survived the emergency situations analyzed their profiles and found a correlation between the patterns in the profiles and the amount of stress recalled.

Summary

The reported studies examined the effects of a variety of psychological conditions on speech, ranging from stress induced by cognitive effort to strong emotions induced by intense psychological and physiological pressures. Although individual variations were observed in all studies, particularly those involving laboratory-induced stress, certain general observations can be made. First, individual psychological conditions clearly affect the quality of speech production. In general, F0 contours for fear and anger are higher and show wider fluctuations than for a neutral condition. Speaking rates tend to increase, accompanied by fewer pauses, in the presence of fear or anger. Other psychological states affect pauses differently,

however. More frequent and longer pauses are likely to occur in a sorrowful, or depressed condition than in neutral. The length and location of pauses also differ for tasks requiring cognitive planning; they are less likely to occur at grammatical junctures and tend to be longer.

Intense emotions such as those occurring in life-or-death situations are unlikely to occur in conditions of typical computer usage. However, sufficient evidence from the neurological, psychological, educational, and HCI literature suggests that the type of task addressed in the proposed research will produce a cognitive burden in addition to the normal workload required for any computer interaction. Emotions likely to be experienced in such situations are gradations on the spectrum of fear and anger, such as anxiety and tension. The level of individual variations observed in the studies of laboratory-induced stress, however, suggest that some users may react instead with feelings of defeat, manifested as gradations of depression or sorrow. Still others may exhibit no reaction which can be detected in the voice in any way. Even if the user experiences no feelings of frustration, Goldman-Eisler's research (1968) suggests that the cognitive planning required for the task may affect the presence of pauses in speech. How might all of these observed changes affect the development of algorithms for automatic prosodic detection? Variations in F0 contours could adversely impact the performance of phrase or tune recognition as well as prominence detection algorithms based solely on this feature. Variations in speaking rate and pauses, however, could also negatively impact the performance of algorithms which include such features as pauses or duration lengthening in phrase or prominence detection. The

importance of certain features in detection may also change under these conditions, e.g., pauses may be less strong a predictor of phrase boundaries than otherwise, likewise for F0 features. The possible variability in the expression of all these features in the speaker prosodics suggests that the use of multiple acoustic features would be required for optimum performance of prosodic detection algorithms.

CHAPTER III

EXPERIMENTAL DESIGN

It is the hypothesis of this research that the prosodic patterns of user speech produced to access spatial information through a displayless interface, employing only spoken language as the input and output modality, differ significantly from those of user speech produced when the interface employs an additional output modality. The hypothesis is assumed to apply to all users, regardless of visual capability; however, users with visual impairments, who are more experienced in coping without vision, are expected to behave differently in the absence of a visual display than sighted users. Thus, testing the hypothesis requires the participation of users with and without visual impairments.

Experimental Testing

General Approach

The hypothesis was tested by analyzing recordings of user speech interactions with a prototype displayless interface to a GIS database available to the U.S.Army Corps of Engineers (USACE) Waterways Experiment Station (WES). The GIS contained data with spatial relationships of sufficient complexity to test the hypothesis. One experiment, consisting of two sessions, was conducted. In the first session,

subjects were tested using a purely displayless interface with no other output modalities, visual or otherwise. In the second session, a tactile interaction device was used to augment spoken language presentation of the data to subjects with sight loss; sighted subjects viewed a graphical display of the data in addition to the spoken language output.

Experimental Treatments

In both sessions, a series of overall tasks, each consisting of a set of subtasks, was given to the subjects. In order to accomplish the overall tasks and subtasks, the subjects queried the database for information. For example, an overall task entailed determining a route, on foot, from location A to location B in one geographical area. Subtasks included locating streets running in the direction A-B. For streets located, sidewalks or footpaths required identification, as well as intersecting streets or any other obstacles in the path. The overall tasks were presented in series of four, with each task in the series containing increasing spatial complexity. The spatial complexity of the tasks was varied in order to obtain additional information on how the spatial aspect of the data affected the results.

Spatial Complexity of Tasks

The design of the spatial complexity of the tasks was based on techniques used by specialists in the field of Orientation and Mobility (O&M) for persons with visual impairments (Jacobson 1993). Four basic route patterns were employed. The patterns, listed from most simple to most complex, are named by the letters of the alphabet

which most closely describe their shape, I, L, U, and Z. An I-route consists of a straight line segment and places the traveler facing in the same direction at the end of the route as at the beginning. An L-route contains one lateral turn and leaves the traveler facing approximately 90 degrees to the left or the right of the original direction. Both U- and Z-routes contain two lateral turns; however, the U-route leaves the traveler facing the opposite compass direction from the beginning after completing the route, while the Z-route leaves the traveler facing the same compass direction as the beginning after completing the route. The four basic patterns are illustrated in Figure 1.

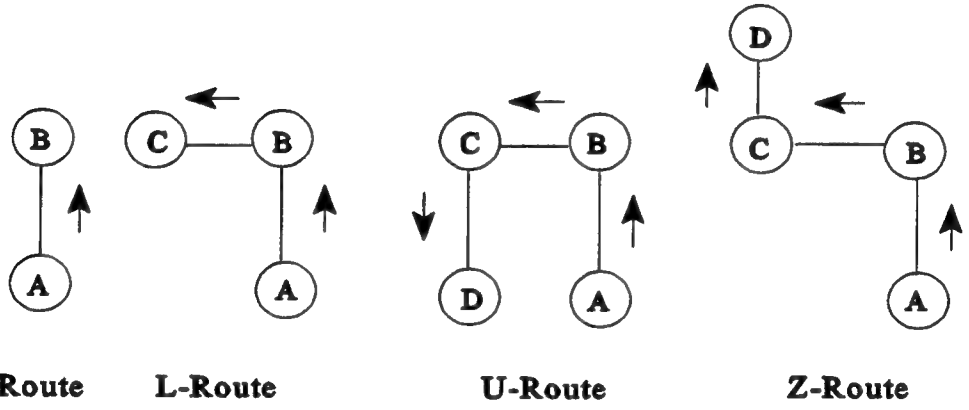


Figure 1. Basic Route Patterns

While these basic patterns formed the foundation for the four tasks, additional factors, i.e., the number and type of required street crossings, increased complexity within each route. Three basic types of street crossings were possible, including

vertical, horizontal, and diagonal. Figure 2 illustrates the types of possible street crossings in order of complexity from most simple to most complex.

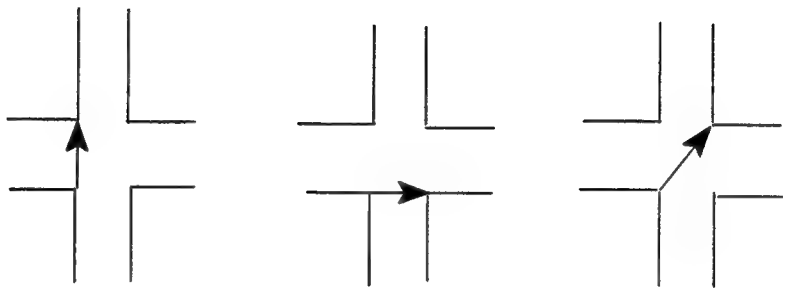


Figure 2. Possible Street Crossings

It is important to note that an arc A-B in the basic route pattern can contain more than one road segment. (The road and traffic network is described in more detail in Chapter 4.) Thus, various road and traffic conditions could force the traveler to choose street crossings within any arc of any route. Consider the example in Figure 3. The route from Environmental Laboratory (EL) to Headquarters (HQ) constitutes an I-route since the traveler essentially follows a straight line and completes the route facing in the same direction as at the beginning of the route. However, three road-traffic segments and thus, three possible street crossings are contained in this route. The first segment, labeled "1", extends from the EL entrance, marked with an "X", to a 4-way

intersection; the second segment, labeled "2", extends from the first 4-way intersection to a second 4-way intersection just prior to the final segment. The final segment extends from the second 4-way intersection to the HQ entrance, also marked with an "X". The traveler's path is shown by a dashed line. The diagram is also labeled to show that road conditions on segment 2, i.e., heavy traffic and no sidewalk, on the right side of the road require a street crossing. Notice however, that the traveler continues throughout the route in the same general direction toward the destination and ends the route pointed in the same direction as at the beginning of the route.

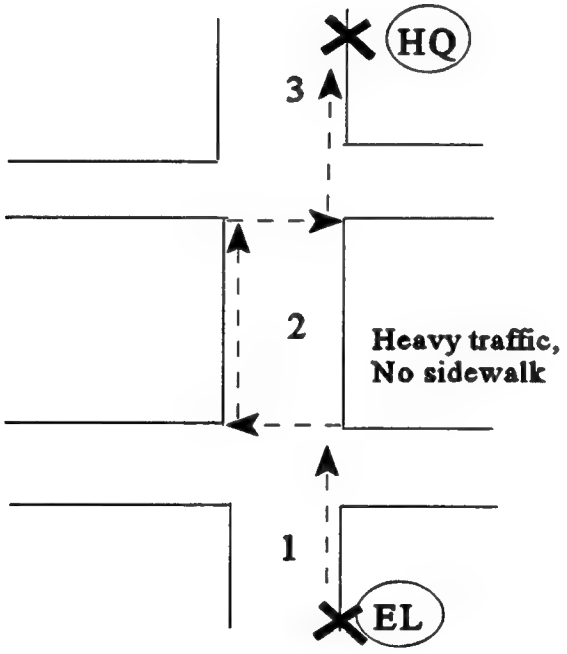


Figure 3. I-Route with Street Crossings

Each of the four tasks in the series corresponded to the basic route patterns, I, L, U, and Z, respectively. A maximum of four street crossings for each arc in a given route pattern was permitted. For instance, an I-route contains one arc, thus a maximum of four possible street crossings; an L-route contains two arcs, thus a maximum of four per arc or eight total street crossings per route, etc. In addition, I-routes and U-routes contained only vertical and horizontal crossings while U- and Z-routes contained diagonal crossings as well. Finally, for the final task in the series, a Z-route, the spatial layout became dynamic. In other words, features in the scenario changed during the route, e.g., a road closed due to an accident.

Subject Preparation for Testing

Before beginning the experiment, subjects were given a verbal description of the nature of the tasks they would perform as well as a description of the spatial layout of the area in which they would perform the tasks. A copy of the descriptions read to the subjects is included in Appendix A. Subjects were given approximately 45 minutes to complete each session with a break between sessions of approximately 10 minutes.

Since a relatively large number of subjects were needed for the experiment, i.e., approximately 90 subjects, training a speech recognizer for each subject was not feasible. (This issue is discussed in greater detail in the following chapter.) Also, the nature of the tasks was of sufficient generality such that no special expertise in a particular application was necessary. In addition, it was anticipated that the levels of experience with a GIS among subjects, both with and without sight loss, would likely

vary and could introduce bias. However, the use of natural spoken language as input eliminated the requirement for knowledge of a specific GIS query language; thus, no special training in the application prior to the experiment was necessary.

Nonetheless, subjects were given time to perform an abbreviated task prior to beginning the experiment as a "warm-up" session to reduce any effects of nervousness or performance anxiety. The task entailed entering a starting and ending point for a route and planning the first two road segments of the route. Subjects were not given a strict time limit on this task, but most subjects completed the task in approximately 10 minutes.

Subjects

General Criteria

Several criteria were considered in selecting subjects for the experiments. A central assumption of the research is the difficulty of nonvisual access to spatial data for all users; however, users with visual impairments were expected to behave differently. Hence, users with and without visual impairments were tested. Since the conditions in the second session differed for the two groups, i.e., subjects with visual impairments employed a tactile interaction device while sighted subjects employed a visual device, no formal statistical comparisons could be made between the two groups. Therefore, data was gathered and labeled separately for each group. Certain criteria applied, however, to all subjects, regardless of visual capability. These include age, education, and amount of prior computer usage. All subjects were required to be of adult age

(i.e., at least 18 years of age or above), possessing the equivalent of at least a high school education (i.e., high school diploma or General Equivalency Diploma). In addition, all subjects were required to be current users of computer software who perform some task on a regular, i.e., weekly or monthly basis. There were no restrictions on the type of software or the task. This was to ensure a minimum level of experience and comfort in computer usage. While the subjects did not need extensive, specialized computer expertise, a complete lack of experience in using computers would clearly impact the difficulty of the task for those subjects.

Criteria for Subjects with Visual Impairments

To restate, although the research hypothesis was assumed to apply to all users, those with and without visual impairments were considered separately. Several issues regarding subjects with vision loss were addressed. First, computer users with visual impairments are typically more experienced in the use of synthesized speech; hence, control over the reading rate of the synthesized speech was provided. More generally, it was anticipated that these users might be better accustomed to the stresses of coping without vision than their sighted counterparts. This would likely differ, however, for subjects with adventitious versus congenital sight loss. The presence of visual memory for subjects with adventitious sight loss was considered likely to impact the results of the experiments. Therefore, subjects with congenital sight loss were considered separately from those with adventitious sight loss. Subjects indicated the time of onset of vision loss in a questionnaire, included in Appendix B.

Several other issues were addressed with regard to subjects with visual impairments. Since the interface used in the second experiment employs a tactile display, subjects were required to possess sufficient tactile sensitivity to enable them to use a tactile interface. This precluded some potential subjects, particularly those with neuropathy in the hands and fingers. Also, consistency in the level of visual impairment of the subjects was required. Individuals who normally relied exclusively on a computer screen magnification program were not deemed the best candidates since their tendency to rely on vision could interfere with their acceptance and use of the tactile display as well as the speech output. Therefore, this category of users was avoided in the experiments.

Criteria for Sighted Subjects

Certain issues regarding sighted subjects were also considered. These subjects, because they possess visual memory, could possibly be better at mentally constructing visual representations of the problem and data than subjects with visual impairments. (Again, this could be less true for subjects with adventitious sight loss.) However, sighted users who employ a more visually-oriented cognitive style may rely more heavily on the use of vision, and thus experience greater handicap in the absence of a visual display. Therefore, a questionnaire designed by Paivio and Harshman (1983) was administered after the experiment to ascertain a dominant cognitive style, whether visual and spatial or verbal and acoustic. The questionnaire is included in Appendix B. This information was used in the interpretation of the experimental results.

Data Analysis

User Data: Prosodics

Prosodic Variables

After completion of the experiment, the recordings for each session, i.e., displayless session and multimodal session, for each subject, were transcribed and hand-labeled, using the Tones and Break Indices (TOBI) transcription system (Silverman et al. 1992). During this post-processing phase, acoustic data for the following prosodic variables was extracted and labeled per utterance: pauses (type, quantity, and length in seconds), breaths (quantity and location), intonational boundary tones (type and quantity), duration (in seconds), preboundary lengthening (in seconds), speaking rate (in seconds), and disfluencies (quantity). Acoustic data for each of these variables was extracted and measured per utterance, where an utterance is defined as a spoken query by the user, or more specifically, a verbalization (which can be a word, partial word, group of words, or group of partial words) spoken by the user to the system in anticipation of a response or a single set of responses from the system. The per-utterance measurements of the prosodic variables were averaged per session as well as per task for statistical analysis.

Statistical Analysis of Prosodic Variables

To determine statistically significant differences in the prosodic variables measured in the displayless session versus the multimodal session, a matched-pair t test was performed. The test compared the means of the differences in the prosodic

variables measured in the displayless session against the prosodic variables measured in the multimodal session. A matched-pair t test (Dowdy and Wearden 1983) was used since the experiment examined two matched groups; i.e., the same subjects were given a "before" treatment (a single, verbal output modality) and an "after" treatment (an additional output modality). These tests were performed for both overall session-to-session comparisons as well as task-level comparisons, i.e., matched-pair t tests were performed for each subject category, comparing the prosodic data for all tasks completed in the displayless sessions against the prosodic data for all tasks completed in the multimodal sessions. Final tests were performed on a task-level basis, i.e., prosodic data for the first task in the displayless session was compared to prosodic data for the first task in the multimodal session; likewise for each subsequent task.

System Data: Speech Recognition Errors

Categories of Recognition Errors

Recognition errors present perhaps the most significant disadvantage of speech interfaces. Not only are recognition errors frustrating for the user, they make it difficult for the user to develop clear mental models of the system's behavior. A variety of phenomenon contribute to misrecognition errors, including speaking before the recognition component is prepared to receive input, false starts in speech, background noise, and out-of-vocabulary utterances, among others. Schmandt (1994) identified three types of errors which can occur in a speech interface, rejection errors, substitution errors, and insertion errors.

Rejection errors occur when the recognizer cannot identify the input or form any hypothesis about it. This can be caused by background noise, early starts in speech, or false starts followed by self-repair. Early starts occur when the user begins speaking before the recognition component is ready. False starts occur when the speaker begins to speak, makes an error, realizes the error, and attempts to correct it. Simply allowing the user to repeat the command can often rectify this problem as well as early starts if they occur (Yankelovich, Levow, and Marx 1995). Also, on repeated rejection errors, the user may begin to speak in an exaggerated tone, making recognition by the system even more difficult. Simply reminding the user to speak clearly and normally can help rectify this problem. Finally, out-of-vocabulary utterances may cause rejection errors.

Substitution errors occur when the system misinterprets the speaker's input and "substitutes" an incorrect interpretation for the correct one. While rejection errors can cause frustration, these types of errors can have more destructive consequences. Kamm (1994) and others (Stifelman et al. 1993;Yankelovich, Levow, and Marx 1995) believe that confirmation of utterances should be commensurate with the cost of an error and that confirmation of every request should only occur if the consequences are critical. Since the primary consequence for the experiments would be frustration for the user and confirmation of every request may tend to increase user frustration, a much more limited confirmation strategy was applied. Confirmation only took place if an error could significantly increase the user's time on a task.

Insertion errors occur when the system attempts to process something which was not a speaker utterance. This is typically due to background noise. The experiments were performed in an office or laboratory, which can be surprisingly noisy due to air conditioning, computers, and peripherals. Since little can be done to alter these conditions, the best method for handling the errors which they cause is to allow the speaker to reenter the input. Some background noise can be introduced if the speaker momentarily is interrupted from the task, intentionally or otherwise. This was addressed by providing the speaker a means of turning the voice input off when not directly engaged in the task of entering data as well as a simple means of reactivating it. Yankelovich, Levow, and Marx (1995) recommend a simple keypad press to deactivate the voice and reactivate it, followed by a spoken prompt from the system when it is ready to begin processing again. A similar strategy was provided for the experiment.

The kinds of recognition errors which occur as well as the system strategies for handling them can impact the level of emotional tension experienced by the users and vice versa. Therefore, during each session, the number and types of errors made by the system were analyzed.

Experimental Measurements of Recognition Errors

Recognition errors made by the system on utterances spoken by the user were calculated per utterance and then averaged per session as well as per task. Each digitized utterance was recorded and saved with a corresponding file containing the text representing how the system recognized the utterance. The words for the digitized

speech were hand-labeled during post-processing. For example, assume the utterance, "Is there a sidewalk?," is recognized by the system as "Is there a landmark?". A digitized speech file containing the original utterance was recorded and saved during the session, along with a corresponding text file containing the text, "Is there a landmark?". The digitized speech was then labeled during post-processing, so that the information in the text file could then be checked against the labeled digitized speech file.

Recognition error analysis was performed on a semantic, rather than a word-level basis. This strategy was chosen since the application used in the experiment did not require the accuracy of a dictation-style program. Since it functioned primarily as a database query language processor, correct recognition of the meaning of the user's request was considered an accurate response. This was particularly important when calculating substitution errors. Consider the example utterance, "How about the sidewalks here?," recognized as "What about sidewalks here?". Although the system substituted "How" for "what", it recognized enough of the utterance to provide a response which would satisfy the user's request. Therefore, such a misrecognition was not counted as a substitution error in the analysis. However, consider the previous example utterance, "Is there a landmark?," recognized as "Is there a sidewalk?". Such a misrecognition is counted as a substitution error since the system would respond with information concerning landmarks rather than sidewalks, which the user requested. Multiple substitution errors per utterance were possible. However, the same criteria

which applied for single errors applied for additional errors, i.e., only semantic errors were considered.

A semantic approach to error calculation was also applied for insertion errors. These types of errors differed from substitution errors, however, in that utterances with insertion errors contained "filled pauses", i.e., pauses containing non-verbal noises such as false starts, background noise, etc., which were labeled during post-processing. For example, "Is there (laughter) a sidewalk here?," recognized as "Is there a landmark or a sidewalk here?," is counted as an insertion error. The background noise due to laughter, labeled during post-processing, is recognized as "landmark," causing a response that does not correspond to what the user requested. However, "Is there (laughter) a sidewalk here?," recognized as "Where is a sidewalk here?," is not counted as an insertion error since the system would respond appropriately to the user's request. Multiple insertion errors per utterance were possible. Again, only semantic errors were considered.

A single criteria was used in counting rejection errors. A text file containing no hypothesis from the speech recognition system regarding the user's utterance signified a rejection error. Only one rejection error per utterance was possible.

Statistical Analysis of Recognition Errors

Statistical analysis of recognition errors made by the system was conducted in the same manner as that for the prosodic variables since the same experimental conditions were applied. To restate, a matched-pair t test was performed comparing

the means of the differences in the measurements of recognition errors extracted from the displayless session versus the multimodal session. Again, a matched-pair t test was used since the experiments examined two matched groups; i.e., the same subjects were given a "before" treatment (a single, verbal output modality) and an "after" treatment (an additional output modality). These tests were performed for both overall session-to-session comparisons as well as task-level comparisons, i.e., matched-pair t tests were performed for each subject category, comparing the recognition errors made by the system on utterances spoken by the user for all tasks completed in the displayless sessions against those for all tasks completed in the multimodal sessions. Final tests were performed on a task-level basis, i.e., recognition errors made by the system on utterances spoken by the user for the first task in the displayless session were compared to those for the first task in the multimodal session; likewise for each subsequent task.

Scope of Study

Certain restrictions on the scope of the research were identified prior to conducting the experiment. First, computer users with visual impairments constitute a low incidence population, which restricted the process of subject selection. Further, obtaining sufficient numbers of both subjects with congenital vision loss and subjects with adventitious vision loss increased the difficulty of subject recruitment. Therefore, sample sizes for each category were anticipated to be relatively small, with a goal of approximately 30 subjects per category.

Also, as mentioned in the description of the subject criteria, only those subjects with vision loss who rely on the use of synthesized speech for computer usage were included in the study. This eliminated the participation of an entire category of computer users who are considered "low vision" users and rely on screen magnification for computer usage. Thus, no data could be gathered on this category of users. In addition, only subjects possessing sufficient tactile sensitivity to use the map in the multimodal session were included in the study, which excluded a category of users with neuropathy in the hands and fingertips. Nonetheless, those users with vision loss included in the research, i.e., digitized speech users with tactile sensitivity, constitute a population of significant size for investigation.

Finally, the experiment required subjects to participate in a single session using a displayless interface followed by a single session using a multimodal interface, thus focusing on short-term, i.e., one-time, users of displayless interfaces rather than long-term users. This precluded the possibility of obtaining information on how subjects would adapt to the interface with repeated use. However, since many displayless applications entail only one-time usage, short-term users comprise a population of significant size for examination.

This chapter described the general experimental design employed in the research. Before presenting an analysis of the results, however, it is necessary to give a more detailed description of the instruments and procedures used in the experiments, i.e., the speech-based prototype and the prosodic labeling scheme. The following chapter provides this description.

CHAPTER IV

EXPERIMENTAL PROCEDURES

Speech-Based Prototype Navigational System

Overview

As stated, the prototype used in the experiments offers speech access to a USACE WES database. More specifically, the database represents the WES Masterplan, which contains details of the spatial and physical layout of the station. The speech interface provides access to the program, "WES Auto Travel," which uses the information in the database to assist first-time visitors in navigating the station via spoken instructions. In both experiments, subjects were asked to play the role of first-time visitors to the station and to use the program for assistance in getting from one location on the station to another. After hearing a verbal description of the overall station layout, given in Appendix A, subjects were given a starting point and destination for each task and then asked to use the program to determine how to get to the destination. Although the program defaults to giving instructions along a precomputed driving route through the station, in both experiments, subjects were asked to customize the route for walking by issuing various queries and commands.

In the first experiment, subjects used a purely displayless interface with no additional modalities, visual or otherwise. All interactions between the user and the

computer took place via spoken language as shown in Figure 4. Also shown in Figure 4 are the hardware and software modules comprising the system.

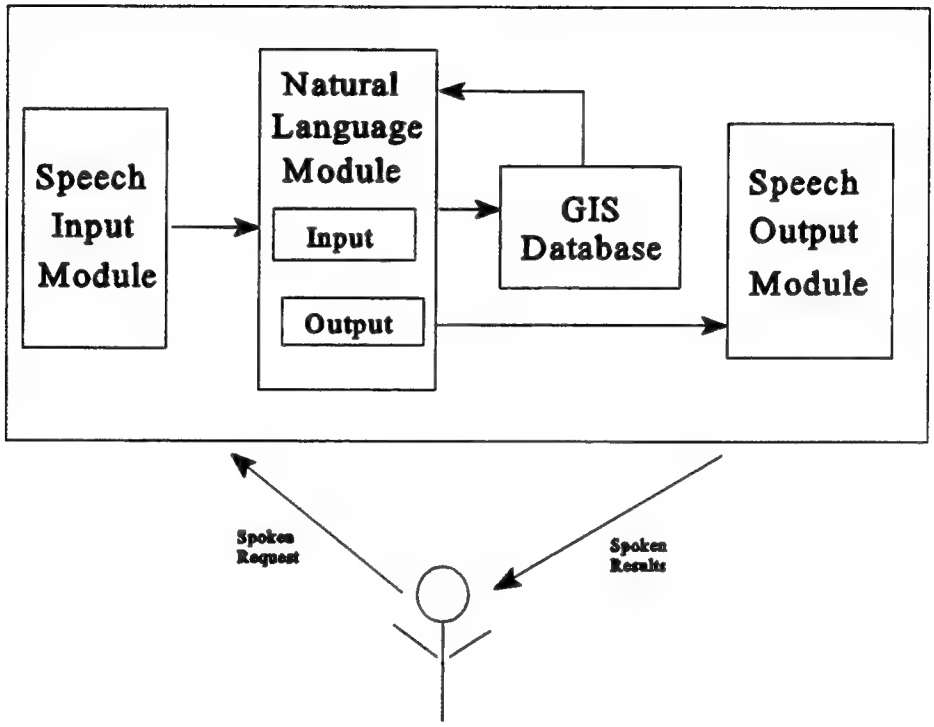


Figure 4. Displayless Access to Prototype

To summarize the flow of control between the individual modules, the Speech Input Module receives the spoken request from the user, processes the speech signal through several steps, and produces a string of words. This word string is then passed to the Natural Language Input Module, which parses the request, translates it into a database query, and presents the query to the GIS database. The result of the query is then returned to the Natural Language Output Module, which formulates the result in natural language and passes a response to the Speech Output Module. Finally, this module presents the natural language response to the user via synthesized speech.

In the second experiment, subjects were given multimodal access to the prototype via a touch screen visual display of a map of the station. Certain key areas or "hot spots" were identified as selectable on the map. The map was represented visually for sighted users and tactilely for users with sight loss; thus the selectable areas were highlighted with visual and tactile markings respectively. Pictures of the visual and tactile map are included in Appendix C. Users could touch the selectable areas on the map and receive auditory information in addition to querying for other information via speech. The flow of information for multimodal access is shown in Figure 5.

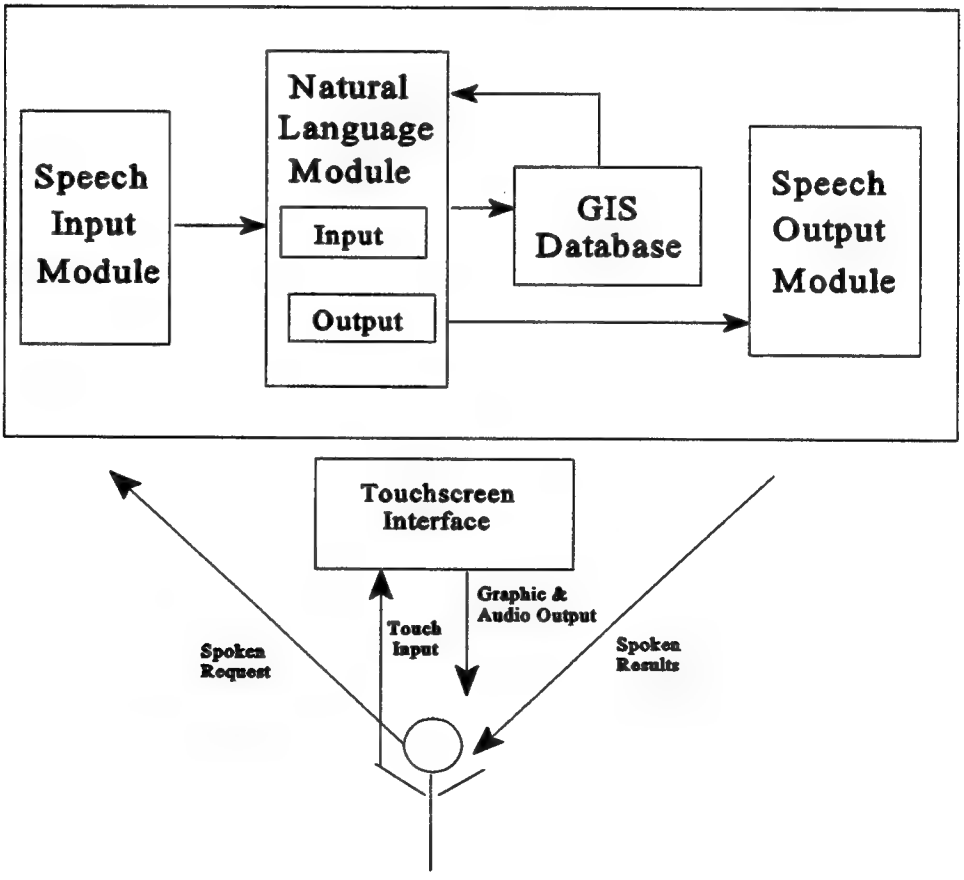


Figure 5. Multimodal Access to Prototype

The modules were developed iteratively and are presented in an order similar to that in which they were developed, e.g., iterative development cycles between the GIS module and the NLP module were followed by iterative cycles between the NLP and Speech module, which were followed by reiterations of the entire cycle.

GIS Database Module

Design

The WES Masterplan database was selected for the prototype application because it offered sufficient, yet manageable spatial complexity for the purposes of the research. Its initial design was influenced by research which examined the use of spoken directions for navigational assistance (Streeter, Vitello, and Wonsiewicz 1985; Davis and Schmandt 1989). These studies provided guidelines for developing a base system which could then be tested and refined.

Data Design

From the outset, it was clear that some type of network data structure involving a set of nodes and directed edges would be needed by the navigational program. However, research described in (Streeter, Vitello, and Wonsiewicz 1985) and (Davis and Schmandt 1989) demonstrated the importance of making this structure correspond to user's intuitions and assumptions regarding traffic patterns and basic navigational operations. For example, people naturally speak of a "Y" or "fork" at an intersection when giving directions. Thus, a data structure representing this type of intersection would be needed. More generally, representing the data internally in ways that

matched people's perceptions would simplify the later task of speaking instructions intuitively to users. In keeping with this goal, a basic network structure was defined, implemented, and then iteratively refined through a series of user tests. Examples of node and edge types used in the final network are described in the following paragraphs.

Road network nodes. Two classes of nodes are represented in the network, traffic nodes and landmark nodes. Traffic nodes include any object or occurrence which significantly alters the flow of traffic, whether vehicular or pedestrian. Examples include intersections, (two-, three-, and four-way, with or without stop signs), forced turns (left and right), sharp curves or turnarounds, and sidewalk or crosswalk endpoints. Landmark nodes are represented in three subclasses. The first subclass contains places at which people gather, including buildings or sites which can be visited on the station, such as laboratories, office buildings, and tourist stops. The second subclass includes objects which are part of the physical infrastructure of the station, such as parking lots, bridges, guardrails, and traffic signs. The third subclass includes naturally occurring landmarks adjacent to other nodes or edges in the network, e.g., a lake running under a major road segment or a wooded ravine directly adjacent to a road segment or footpath.

Road network edges. Two classes of edges are represented in the network, vehicular and pedestrian. Vehicular edges are simply segments of road upon which motor vehicles can travel. This class contains several subclasses, including road segments with narrow or no shoulder, road segments with wide shoulder, road

segments with adjacent sidewalk, and road segments with no adjacent sidewalk. Pedestrian edges, e.g., sidewalks, crosswalks, and footpaths, are represented only when connected to a vehicular edge, landmark node, or traffic node in the network. A sidewalk connecting two office buildings (landmark nodes) provides an example of a pedestrian edge that is represented in the network. Conversely, a footpath in an open field unconnected to any other nodes or edges in the network would not be represented. However, if the same footpath or field were adjacent or otherwise connected to a vehicular road segment (vehicular edge), a tourist stop (landmark node), or a crosswalk endpoint (traffic node), it would be represented. Two issues motivated the selection of pedestrian edges, availability in the database, i.e., the open field in the previous example is not likely represented, and human factors, i.e., entities unconnected to other objects in the database are likely more difficult to describe, hence more difficult for the user to locate.

Finally, all edges are stored with an associated direction as well as additional information about the edge. This includes such data as the level of traffic and average speed of vehicles on the edge if it is vehicular as well as pointers to adjacent landmarks along the edge if it is vehicular or pedestrian.

Application Design

The application was designed to allow varying levels of interactive control by the user, ranging from offering the user a precomputed route to allowing the user to fully construct a route with only passive assistance from the routing program. In the

experiments, subjects employed a mixed level of interactive control, taking a precomputed route and customizing it for their individual purposes and preferences. Regardless of the level of interactive control, a user can address the program in one of two modes, command mode or query mode. In command mode, a user can issue commands, for example, to continue to the next segment on the current route, go back to the previous segment, or calculate a new route beginning at the user's current location. In query mode, a user can ask questions about the current route or the station in general. Examples include the speeds of cars and level of traffic on a particular road segment, landmarks in a specified vicinity, orientation with respect to a major landmark, distance to a destination, etc. The linguistic aspect of these features was addressed in detail in the development of the Natural Language Module and is discussed further in the section describing that module.

Implementation

A Sun UltraSparc workstation served as the development platform for the prototype. Several factors contributed to this choice of development environment. First, the native audio as well as the general computational capabilities of the Sun satisfied the interactive requirements of the prototype. A second determining factor concerned availability of software. In addition to supporting a large selection of speech recognition research tools, this environment was compatible with that of an existing Environmental Systems Research Institute (ESRI) ARC/INFO GIS database, containing the WES Masterplan (ESRI 1995). Other commercial software used in

construction of the prototype included HTK speech recognition software, produced by Entropic Research Laboratories (Entropic 1996), used in the Speech Input Module, and Centigram Truvoice speech synthesis software for the Speech Output Module (Centigram 1996). Further details are given on the use of the latter tools in the sections detailing those modules.

As stated, an ESRI ARC/INFO database provided the relevant GIS data, which was extracted and translated into flat files containing the network structure described previously. The routing software, developed in the C++ programming language, accesses these file structures to produce an optimum route or path between two given nodes in the network. Several criteria define the optimality of a path. These include physical distance as well as complexity, e.g., number of segments in the path, number of turns, forks, etc. Thus, a path shortest in physical distance will not necessarily be chosen as optimal, particularly if its complexity is high, containing numerous segments and difficult turns. Such a route, though shortest physically, would likely be more difficult to follow, hence increases the cognitive load on the user.

In addition to the basic routing software, a suite of ancillary functions implements the possible queries users can present to the system. This suite includes functions for basic orientation queries, e.g., determining the user's current location relative to major landmarks or a destination, as well as other types of queries, such as determining speeds of cars and traffic levels on road segments and locating sidewalks, crosswalks, or landmarks.

Two techniques were implemented to resolve landmark queries, depending on the context of an individual query. The first assumes the user is stationary, in which case only those landmarks in the user's immediate vicinity are presented. The second assumes the user is moving along a route, in which case landmarks are presented in the order in which they would be viewed traveling from the current segment in the route to the next segment.

This concludes the description of the GIS Module. The design and development of this module and that of the Natural Language Module were closely linked; the next section describes the latter module.

Natural Language Module

Design

Although implementation of this module began only after a rudimentary version of the GIS module was constructed, design of the two modules proceeded in tandem. Again, the research of (Streeter, Vitello, and Wonsiewicz 1985) and (Davis and Schmandt 1989) emphasized the importance of understanding not only the users' intuitions about navigational problems, but how they express these intuitions in words. A technique referred to in the speech recognition literature as the "Wizard of Oz" technique (Honma and Nakatsu 1987) provided a tool for quickly eliciting this information.

Simply described, this technique enables a user to interact in an electronic session with a mock application controlled by the interface or application developer.

The user is allowed to pose requests using whatever sequence of words or phrases comes to mind. The developer or "wizard" interprets the request. If it is a request which the application can compute and satisfy, regardless of how it is phrased, the "wizard" presents the results of the query to the user. A transcript of the session is saved for later analysis.

Twelve individuals employed in the WES Information Technology Laboratory (ITL) participated in a series of such sessions, conducted in an Electronic Meeting Systems (EMS) environment over the course of two months. The individuals were selected to capture a diversity of possible occupational, cultural, and gender-based biases. In addition, efforts were made to involve individuals with physical disabilities, including one person with a visual impairment. Although these sessions are typically implemented with a surrogate speech recognition component, the EMS environment did not allow for an implementation of such a surrogate convincingly. Therefore, participants in the WES sessions either entered their responses via keyboard or had a typist enter their responses while they spoke.

The sessions yielded several benefits, without requiring implementation of all possible user inputs. First, they provided a record of words and phrases users would speak, given no constraints. Information gleaned from this record was used to enhance the coverage of the vocabulary and grammar, thus making them more robust. For example, the sessions demonstrated that people often used the word "bad" in relation to "traffic," but this could refer to either the volume or speed of the traffic. In particular, when crossing at an intersection with no crosswalk or traveling around a sharp curve,

high speed traffic, even if occasional, could present more danger than large volume, but slow, constant and predictable traffic. Thus, "bad" was included in the grammar as a modifier for a traffic attribute, but represented as one that might require clarification from the user, depending on its context. Another frequently used word, "condition," as in "condition of the road" or "road condition" presented similar problems. Users spoke these phrases, at times, in referring to traffic conditions, in which case they desired to know the volume and speed of traffic. However, this expression could also refer to the actual physical condition of the road, whether it is paved or unpaved, hilly, curvy, or if there is any construction on the road, etc. Again, this phrase was included in the grammar, but represented as one requiring possible clarification from the context.

Second, these sessions offered insight on the preferred methods for presenting information to the user. This included how the information should be phrased, as well as how much information to provide and when. For example, when the program presents a precomputed route, users preferred that both the physical distance of the route and the number of segments it contains be spoken at the beginning of the route. Thereafter, however, they desired the physical distance of individual segments to be spoken only if the segment were particularly long or short.

Presentation of directional information for orientation purposes provided another example of the importance of proper phrasing of output. Several studies have indicated that people vary widely in their understanding and use of compass directions, i.e., north, south, east, west, etc. (Kozlowski and Bryant 1977; Thorndyke and Stasz 1980). This proved true in the wizard sessions, which demonstrated the importance of

offering as many kinds of information as possible when giving directions. This entailed using compass directions, commonly used directional language, such as "left," "right," "continue," and prominent stationary physical landmarks. Consider the example, "You are currently located on Hudson Road at the Geotechnical Laboratory. Facing Gate 5, headed in the northeast direction, with the Geotechnical Lab main office building on your left, and the Geotechnical Lab parking lot on your right, continue on Hudson to the intersection of Hudson and Porter's Chapel." This instruction uses compass directions ("northeast"), commonly used directional words ("left," "right," "continue," "facing"), and prominent stationary landmarks (Geotechnical Laboratory main building, Gate 5, Geotechnical Lab Parking Lot). Presenting multiple categories of information addressed the widest range of user preferences and reduced the ambiguity of any one category presented alone. Even those who considered themselves skilled with compass directions preferred the additional information for verification, particularly when unable to view a map.

Finally, the sessions helped to elicit and refine additional functional requirements for the program. Inclusion of an individual with a visual impairment proved especially beneficial in this regard. Although this individual, heretofore referred to as WES-VIP 1, had received minimal formal O&M training, his experience as a day-to-day foot traveler on the station proved invaluable. For example, the database contained information on the amount of shoulder associated with each road segment, whether wide, narrow, or non-existent. WES-VIP 1 provided guidance on the importance of not only the size, but the quality of the shoulder. For instance, is the

surface level or uneven? What is the surface material? Depending upon other conditions in a scenario, a narrow, level surface covered in grass, if mowed, might be preferable to a wide, uneven gravel surface. Therefore, such attributes were added in the description field for shoulder in the database to be presented to users in relevant queries about the road and its shoulder. Thus, the request, "What about sidewalks on this road or does it have a shoulder?", depending on the location, could be answered, "There is no sidewalk on this section of the road, but there is a wide, level grassy shoulder." Interestingly, during usability tests conducted after this feature was incorporated, sighted users also appreciated this benefit.

WES-VIP 1 also offered insight on the inclusion of landmarks with an auditory component. Most participants in the session traveled the station almost exclusively by automobile, and were thus less attentive, and in some cases, unaware of such landmarks. For instance, the WES Nature Trail and Arboretum was included as a landmark in the database since it is a tourist stop; however, WES-VIP 1 noted that it also attracts many seasonal birds and therefore often provides this auditory cue, useful in orientation. Thus, when the program describes this landmark to the user as being to his or her right, the sound of warbling birds emanating from that direction can serve as a confirmational cue. The many hydraulic shelters on the station provide more permanent examples. Several of these shelters house large hydraulic pumps, whose activity is often audible to nearby pedestrians. Therefore, such attributes were added in the description fields for these objects in the database, so that when these landmarks are described to the user, their auditory component is mentioned, e.g., "On your left,

you will pass Waterways Shelter #3. It houses large hydraulic pumps, which can be heard operating on a daily basis."

Since WES-VIP 1 had received only minimal O&M training, efforts were made to incorporate more of this type of knowledge in the system design. Input was solicited from two additional users with visual impairments, WES-VIP 2, who had received more extensive O&M training in cane travel, and WES-VIP 3, who had received O&M training in the use of a dog guide. Though unable to take part in the local wizard sessions, WES-VIP 2 and WES-VIP 3 participated in several telephone interviews conducted during the course of the sessions. Much of their input confirmed design decisions based on suggestions from other participants, including both WES-VIP 1 and other sighted participants. For example, some cane travelers may use an O&M technique called "shorelining," described in detail in (Jacobson 1993), which involves maintaining a line of travel by following changes in the travel surface with the tip of the traveler's cane. Since this technique is often used with sidewalks, the suggestion of WES-VIP 1 to provide information on the quality of the shoulder when a sidewalk is not present served the needs of cane travelers as well. Also, the shorelining technique presents the problem of allowing travelers to inadvertently veer into parking lots since changes in surface texture can be deceptive. The traveler may then spend some time wandering in the parking lot before realizing the mistake. This issue corroborated the decision, based on input from both WES-VIP 1 and other sighted participants, to include parking lots as a landmark.

In addition to specific suggestions, the participation of WES-VIP 2 and WES-VIP 3 demonstrated the difficulty of defining criteria for an optimal walking route for all users. It was anticipated that users with and without sight loss would use different criteria; however, these sessions demonstrated the variability among individuals with sight loss. For instance, dog guide users can typically walk for extended periods at a fast pace since many dogs can travel at speeds up to five miles per hour. Therefore, these travelers may be more willing to accept a longer, but less complex route, than those without this experience. As another example, WES-VIP 1 preferred road segments bearing the least possible amount of traffic, a preference that could be indulged by those already familiar with the station. However, those less familiar may wish to employ a common O&M technique of using the direction of the sound of traffic to verify direction of travel. If so, these travelers may prefer roads with light to moderate traffic over those with occasional or none. Finally, and more generally, the skill and confidence level among travelers with visual impairments varies widely. The confluence of these factors contributed to the design decision to ask subjects to create their own customized walking route in the experiments.

Implementation

The NL Module managed four primary tasks, parsing input produced by the speech recognizer, translating the input to a database query, resolving the query, and presenting the results of the query to the user in natural language. The components shown in Figure 6 implemented these tasks and are described in the following section.

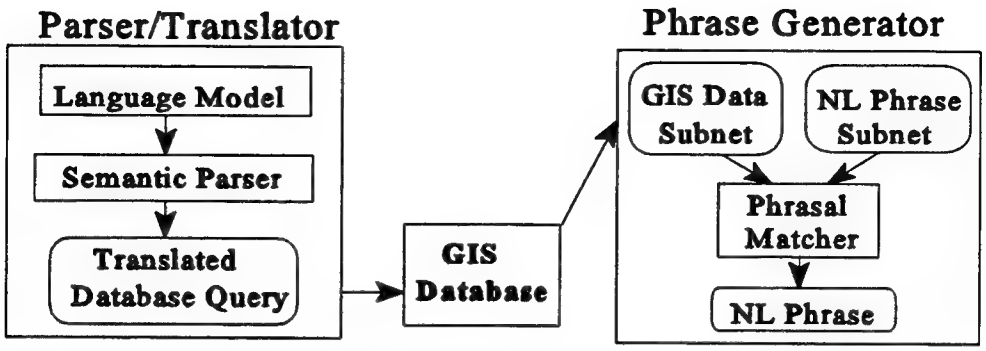


Figure 6. Detailed View of NLP Module.

Parser/Translator

Interspersing the design of the GIS and NL modules proved beneficial in developing a grammar and parser that reflect the functionality of the system. Both the nature of the application as well as the resource constraints on its development made a restricted natural language interface, allowing a mixture of fixed commands and freely formed natural language, most feasible for the application. This decision, in turn, influenced the choice of a semantic grammar for the language model. First used for database applications, (Hendrix et al. 1978; Hendrix 1978), semantic grammars have been shown useful for quickly producing restricted natural language interfaces. This type of grammar uses semantic rather than syntactic categories in the production rules of the grammar. The following simplified fragment, based on production rules in the grammar for "WES Auto Travel," provides an example:

VEHICLE_QUERY -> what is the TRAFFIC_ATTRIBUTE on ROAD?
TRAFFIC_ATTRIBUTE -> traffic| traffic level| speed|speed limit| ...
ROAD -> ROAD_NAMES ROAD_WORDS
ROAD_NAMES -> Osage|Hudson|Porter's Chapel|....
ROAD_WORDS -> road|street|avenue...

Since the grammar uses semantic categories such as ROAD and TRAFFIC_ATTRIBUTE rather than strictly syntactic categories such as NP (noun phrase) and VP (verb phrase), the results of the parse can be used immediately without further semantic post-processing. This decreases the level of processing required for translation to a database query, hence the advantage of this type grammar for use in database applications. Although it presents the disadvantage of requiring larger numbers of rules to capture all possible syntactic generalizations, its benefits for rapidly producing a restricted natural language interface proved more significant for this application.

WES Auto Travel Grammar. As mentioned, the grammar contains a mixture of fixed commands and freely formed natural language requests. Users must issue fixed commands for two general purposes, requesting help or traveling along a precomputed route. For example, when traveling along a precomputed route, the command, "continue," spoken by the user, causes the program to continue forward and provide directions to the next segment of the route. Likewise, the command "go back," causes the program to return or "go back" to the previous segment of the route. Similarly, the command "change route" prompts the routing program to compute an alternate route to the specified destination. Users can request help for remembering these commands at any point along the route, or for general help during any type of interaction with the

system, by issuing the command "What can I ask?". The wizard sessions influenced the initial choice of phrasing for the fixed commands which were later refined during design and testing of the speech interface. More discussion on this aspect of the interface is given in the section detailing the Speech Input Module.

Examples of possible freely formed natural language requests include any type of query to the database, e.g., "How far is it to the Headquarters Building?", "Are there any sidewalks on this section of Mississippi Road?", etc. System functionality imposes the only significant constraint on this type of request. In other words, the request concerning sidewalks on Mississippi can be phrased by the user in numerous ways, "How about sidewalks here?", "Is there a sidewalk on this street?", "Does Mississippi have a sidewalk?", etc., all of which would elicit the system response, "This section of Mississippi has an adjacent sidewalk." However, the request, "How many non-government vehicles traveled on Mississippi last month?" could not be answered since this information is not contained in the database, nor is any query available to access it. Thus, a help command, in this case, "road queries," spoken by the user, prompts the system to provide help on the kinds of information regarding roads contained in the database.

Given the examples of commands and requests available, revisiting the structure of the "WES Auto Travel" grammar shows the following simplified overview:

UTTERANCE -> FIXED COMMAND|GENERAL QUERY
FIXED_COMMAND-> HELP_COMMAND|TRAVEL_COMMAND
HELP_COMMAND-> what can I ask|road queries|...
TRAVEL_COMMAND->continue|go back|....
GENERAL_QUERY-> VEHICLE_QUERY|PEDESTRIAN_QUERY

VEHICLE_QUERY-> what is the TRAFFIC_ATTRIBUTE on ROAD?
TRAFFIC_ATTRIBUTE -> traffic| traffic level| speed|speed limit| ..
PEDESTRIAN_QUERY-> does ROAD have PEDEST_FEATURE?
PEDEST_FEATURE-> sidewalk|crosswalk|footpath|walkway|...
ROAD -> ROAD_NAMES ROAD_WORDS
ROAD_NAMES -> Osage|Hudson|Porter's Chapel|....
ROAD_WORDS -> road|street|avenue...

It should be noted that in addition to semantic knowledge, the program maintains some limited contextual knowledge to aid in query translation. This knowledge consists primarily of a context stack, containing records of previous queries and commands issued by a user in a particular session. This knowledge offset some limitations of the semantic grammar. Referring to the previous query example, "Is there a sidewalk on this street?", since the user does not specify a street name, the translator retrieves the last street name placed on the context stack during the session, i.e., the last street name spoken by the user, in this case "Mississippi," and provides information on sidewalks for that street.

Translation Software. The core software for the parsing module, developed in C++ for a natural language database query system, NLDQS, (Baca and Cooper 1989) was ported to the Sun platform for modification to the "WES Auto Travel" application environment. The grammar and parser contained in the core software were modified to capture the semantic categories inherent in the functionality defined initially in the GIS module. In addition, to restate, insight gained from the Wizard sessions was iteratively incorporated in the modifications. The NLDQS software employed a lexicon of 5000 commonly used words, tagged by part of speech, and a dictionary of proper names specific to the application. Although the full lexicon is not currently used by the

Speech Input Module, it minimally impacts the overhead incurred by the NL module. It was therefore maintained in the NL module with a view toward future enhancements of the Speech Input Module. The NLDQS data dictionary was also modified to contain proper names for the WES application, including building names, road names, etc. Finally, since a semantic grammar was used, the results of the parse required minimal post-processing for the translation to database query, i.e., accessing system contextual knowledge.

Phrase Generator

Similar to the parser, though less complex, the phrase generator reflects the functionality of the application. It employs a natural language phrasal network, constructed to correspond to the network structure produced by the GIS database. It accepts as input a subnetwork of directed edges and nodes from the GIS component. It then uses a phrasal pattern matching algorithm to search the natural language phrasal network and find the appropriate phrasal subnet for the response. An example best illustrates the process.

Consider the query, "What are the landmarks on this road up to the fork in Mississippi?". Assume also that the system contextual knowledge shows the person is currently located at the Environmental Laboratory main building. Given this, the GIS subnet produced by the parser/translator contains a beginning node of class "landmark" and subclass "laboratory," an ending node of class "traffic" and subclass "three-way intersection with no stop sign," (also called "fork") and a vehicular edge (since

Mississippi is a road upon which motor vehicles travel) along with the requested information associated with the edge (adjacent landmarks). The phrase matching algorithm searches the phrasal network, containing words and phrases corresponding to each of the node and edge types in the GIS subnet, to locate and produce the proper phrasal subnet. In this case, the phrasal subnet is shown in Figure 7.

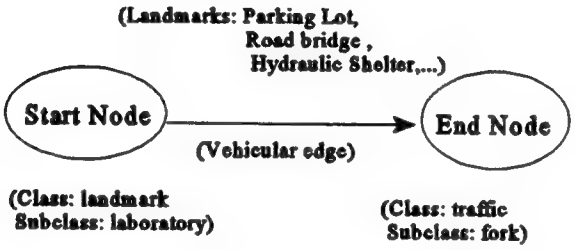


Figure 7. Subnet Produced by Phrasal Generator

The natural language phrase generated by this subnet is: "Starting from your current location at the Environmental Laboratory, continuing towards the fork in Mississippi, you will pass Environmental Lab parking lot on your left. You will then pass over a road bridge above Brown's Lake. The bridge has a footbridge and a guardrail. After the bridge, on your right, you will pass Waterways Shelter #3, a long low, metal building. It houses large hydraulic pumps which can be heard operating on a daily basis."

Lastly, the NL output module generates phrasing for all help text presented to the user. This task, however, does not require use of the phrasal matching algorithm

since these words and phrases are fixed. Further details of this aspect of the interface are given in the section describing the Speech Output Module.

Speech Input Module

Design

Similar to the GIS and NL modules, design and development of the Speech Input Module evolved through successive iterations. Since this module handles the task of recognizing the user's speech and decoding it to a string words which can be parsed by the NL Module, the design of each inevitably influenced the other. However, the research hypothesis dictated the two most critical requirements for the speech recognition component, speaker independence, and recognition of continuous speech.

Speaker independence refers to the capability of the recognizer to understand a large number and variety of speakers without requiring training to a particular speaker's voice. This capability was necessary since approximately 90 speakers would participate in the experiments. Therefore, training to each speaker would not have been feasible. This also increased the flexibility of the interface for later research. The second criteria for the recognizer, the capability to process continuous speech, means that it must allow the user to speak long phrases continuously and naturally without requiring pauses between words. This was essential for the prosodic analysis. The research hypothesis investigates the effects of cognitive stress on the prosodics of the user's speech. Forcing the user to speak in a halting, unnatural manner would clearly impact the results.

Implementation

A commercial speech recognition software toolkit, HTK, produced by Entropic Laboratories, met the critical requirements of the research (Entropic 1996). The Entropic software provides the tools for constructing a large vocabulary, speaker independent, continuous speech recognition application. Before discussing the implementation further, it is necessary to briefly review the basic steps entailed in the recognition process.

A speech recognizer consists of three primary components: 1) a language model for the application, 2) a phonetic dictionary of all words used in the application, and 3) trained statistical models of each subword (phoneme) in the language. The recognizer compiles a network from these components and then attempts to find the most probable path through this network to decode the speaker's utterance. The decoded utterance can be represented as a phoneme string (useful in development stages), but for application interfaces is typically represented as a string of words. For example, the recognizer can output the word string, "Headquarters" or the phonetic string, "hh eh d k w aa rt er z". The latter is useful for locating the source of errors in recognition during development, while the former is typically needed to drive an application.

Implementation of the speech recognizer for "WES Auto Travel" involved assembling each of the three primary components in the HTK environment. Development of the language model and phonetic dictionary proceeded incrementally. A small subset of the semantic grammar and lexicon of the NL Module provided an

initial basis for the language model and phonetic dictionary used in a pilot demonstration in the HTK environment. Translating the semantic grammar to a language model in HTK format proved straightforward. Producing the phonetic dictionary entailed a more extensive process of translating phonetic models taken from the publicly available Carnegie-Mellon University (CMU) triphone pronunciation dictionary to the DARPA phonetic models used by HTK. This process is described in (Ngan and Picone 1997). Once the recognizer functioned properly in the pilot demonstration, the coverages of the language model and dictionary were iteratively increased until deemed sufficient to begin usability testing preliminary to the experiments. The final language model contained over 150 production rules; the final vocabulary contained approximately 200 words.

Usability Testing

A series of user tests conducted prior to the actual experiments served two purposes, revealing design flaws in the speech interface which could affect its usability, and exposing potential problems that could arise in conducting the experiments. Ten individuals employed in the WES ITL, two of whom were involved in the wizard sessions, participated in the tests. Again, these individuals were selected to elicit a diverse mixture of occupations, genders, and regional backgrounds. At the completion of the tests, all participants answered surveys designed to assess the usability of the speech interface. Features of the interface assessed in the survey included intuitiveness of the fixed command language, speed, and accuracy of the recognizer as well as

aspects of the spoken system prompts, which are discussed in further detail in the description of the Speech Output Module.

Command Language. The phrases selected for the fixed commands required some refinement. While the wizard sessions provided a basis for the command language, the usability tests made clear that the limitations of the recognizer would need to be addressed. This required a compromise between what was most intuitive to the user and what the recognizer could best process. First, short, one-syllable words were most often misrecognized. The brevity of these utterances gives the recognizer less contextual information to use in the pattern-match. Some commands were designed in anticipation of this problem, e.g., "what can I ask" served as the command to request general assistance, though users found "help" most intuitive in the wizard sessions. The usability tests revealed other commands requiring modifications as well, such as the word "travel" for requesting help on travel commands, which was replaced with "navigate".

As another related example of command modifications, participants in the wizard sessions most widely preferred to phrase, "Go Forward" to continue forward one segment on a precomputed route. However, in the usability tests, the recognizer consistently misrecognized several speakers when issuing this command due to what is referred to as "coarticulation effects" (Schmandt 1994). In simple terms, this describes a speaker's lack of enunciation of the "f" sound such that the recognizer cannot distinguish the two long "oh" sounds in "Go Forward", interpreting the phrase as something similar to "Go-oward", which would then be matched with the word "road"

in the phonetic dictionary. Substituting the command "Continue Forward" reduced some of the effects of coarticulation. Users also found this an acceptably intuitive substitute for the original phrase. Eventually, per suggestions made in the usability survey, the command was shortened to "Continue".

Recognition Accuracy. Users found the level of accuracy of system recognition acceptable due, in part, to design techniques, but also to adjustments made during usability testing. Refinements to the command language, described previously, significantly reduced the overall error rate, decreasing both substitution and rejection errors. Subsequent to this overall reduction, however, substitution errors continued to occur most frequently in the freely formed natural language queries. Nonetheless, users found these least frustrating since a substitution error does not necessarily result in the program giving incorrect information, particularly in the natural language queries. For example, "Are there any landmarks on this street?" understood by the recognizer as "Where are any landmarks on this street?" could still result in the program providing the information desired by the user.

Allowing users to start and stop the recognizer with a keypress reduced the number of insertion errors, since this reduced the possibilities for background noise or for users to speak before the recognizer was activated. However, some insertion errors were inevitable, for instance if the user coughed, sneezed, or simply misspoke, mid-utterance. Again, when insertion errors did occur, they did not always impede the user in accomplishing a task. In many cases, such errors resulted only in the program presenting additional information beyond what was requested. Consider the utterance,

"Is there high speed traffic on this road?". Assume it is interpreted as "Is there a sidewalk and what about traffic on the road?" This would cause the program to present additional information regarding sidewalks, not requested by the user, but which may also be useful. Finally, users found rejection errors to be most vexing. Since elimination of these as well as the other types of errors was not possible, the error-handling strategies, described below, were designed to limit their negative impact.

Error-handling Strategies. First, as recommended in (Kamm 1994), a minimal confirmation strategy was employed, i.e., the program confirmed the user's request only when the consequences of an error could cause significant inconvenience to the user. For example, when the user requests an alternate route, the program, before calculating a new route and voiding the current one, asks for confirmation from the user by saying, "Your last command was interpreted as a request for an alternate route. If this is correct, press 1 to continue with the operation. If not, press 0 to cancel."

As mentioned, among the errors which could not be corrected through confirmation, users found rejection errors most egregious. Therefore, a strategy similar to that described in (Yankelovich, Levow, and Marx 1995) was employed to handle such errors. This strategy entails rephrasing the program prompts to the user on repeated rejection errors. After the initial unrecognized utterance, "WES Auto Travel" prompts the user, "Could you repeat that please?" If the next utterance is unrecognized, the program prompts, "Could you try that again? Just relax and speak naturally." On the third failed utterance, the program prompts, "Please try rephrasing your request." Finally, on the fourth such error, the program prompts, "WES Auto

Travel is having trouble understanding your request. If you are unsure what you can ask, say 'What can I ask?' and the program will prompt you." This strategy assumes that initially the speaker used words and phrases known by the system, but simply needs to repeat the request due to some temporary acoustic phenomena, such as coughing, exhaling, background noise, etc. The second prompt seeks to ameliorate possible tension or exaggeration in the user's voice due to the first misrecognition. The third prompt addresses the possibility that the speaker may be using words or phrases not in the system vocabulary. The fourth prompt considers that the user may be asking for information not known by the program. This incremental rephrasing accomplishes two purposes. By not repeating the same prompt rotely, it not only seeks to address the cause of the misrecognition, it gives the user the sense that the program is listening and trying to understand, which has a positive psychological effect. This reduces the likelihood that the user will speak the same utterance repeatedly in an increasingly tense and exaggerated tone, causing a downward spiral of misrecognitions followed by increased user frustration.

Response Time. Response time presented the most significant limitation of the speech recognition component. This was anticipated to some degree, since despite several efforts to increase its speed, the recognizer did not run in real time. Latency ranged from approximately 2 to 12 seconds, averaging approximately 5 seconds. Again, several efforts were made to improve the response time. HTK allows developers access to parameters controlling the size of the search space in the recognition network. Simply described, this allows removing or "pruning" from

consideration any nodes in the network whose log probabilities fall more than a certain measure below the token with highest probabilities. HTK allows program manipulation of this measure, called the "pruning beam-width". A larger beam-width increases the search space, resulting in longer search time and consequently, longer response time. Conversely, setting the beam-width too small can cause premature pruning of the correct node in the network, hence an incorrect answer and misrecognition. Therefore, finding the proper beam-width requires a compromise between the speed of the recognizer and its accuracy. Several experiments were conducted for the recognizer used in "WES Auto Travel" in attempts to find the best compromise. Setting the beam-width small enough for the recognizer to run in real time produced an unacceptable number of rejection errors, with an overall error rate of over 60%. Ultimately, the response time was reduced from a minimum of approximately 12 seconds and maximum of 45 seconds (using no pruning) to a minimum of approximately 2 seconds and maximum of 12 seconds, with an overall error rate of approximately 30%.

Experiments were also conducted in reducing the complexity of the language model and vocabulary to determine if this would significantly lower response time. These efforts produced only negligible effects on recognition speed. This could possibly be attributed to the original size of the application; since it is a small to medium vocabulary system, fewer reductions were possible. Regardless of the cause, however, the only remaining alternative to increase the speed entailed additional training of the acoustic models on the application. Such training would involve

conducting wizard sessions, using a surrogate speech recognizer, on a user population equivalent in size and diversity, at a minimum, to that of the subject population in the actual experiments. This was not feasible, given the resource constraints on the prototype development.

It should finally be noted that users did not necessarily find the response time unsatisfactory during usability tests, particularly once the error rate was reduced. Most users had little to no experience with speech recognition applications, thus their expectations on this issue were not rigidly set. (This proved true in the experiments as well.) According to responses in the usability survey, given the choices, "unacceptably slow," "slow, but acceptable," "unnoticeable," "pretty fast," users most often rated the recognizer as "slow, but acceptable".

Speech Output Module

Design

Several criteria determined the design of the Speech Output Module. Many of these have been discussed previously, the most significant of which include seeking to avoid auditory overload, allowing the user control of the volume and speaking rate of the synthesizer, and providing some form of interruptibility. In addition, analogous to the manner in which the design of the Speech Input and Natural Language Input Modules influenced each other, so did that of the Speech Output and Natural Language Output Modules. Similarly, the final design of this module was derived through successive iterations of implementation, user testing, and refinement.

Implementation

Avoiding auditory overload presented a particular challenge for the "WES Auto Travel" program due to the spatial nature of the information it presents. Although the research hypothesis expects the user's cognitive load to be increased by verbal presentation of such data, this could only be accurately tested if auditory overload were first minimized as much as possible. Several measures were taken to achieve this goal, such as keeping prompts as short as possible and avoiding or minimizing the use of auditory lists.

The most crucial, however, concerned the presentation of directional information. Recall that the design sessions for the NL module showed that users preferred multiple categories of directional information. While this reduces ambiguity of the instructions, it increases the amount of information presented to the user and thus, the potential for auditory overload. To minimize this potential, the information is presented into short, related segments. Thus, when a user begins a route, the program gives orientation in several short segments, each repeatable by pressing a key. For instance, the initial segment simply describes the user's current location, giving its orientation to general landmarks, e.g., "You are located at the Headquarters Building. It is situated just inside the main gate to the station on Hall's Ferry Road. You are pointed south, facing the WES Visitor's Center with the main gate to right." The user can then press a keystroke to hear the information again, as many times as desired, until ready to continue. On continuing, the program then describes the destination, "Your

destination, the Geotechnical Laboratory, is located approximately 1.8 miles northeast of your current location. Its nearest major landmark is Gate 5, the northeast entrance to the station." Again, the user can request that the program repeat these instructions until ready to begin the route. Similarly, once the user begins the route, the program presents directions along the route one segment at a time, allowing the user the opportunity to query, give commands or request repetition of the instructions before continuing to the next segment. Other efforts were made to reduce information overload along the route as well. For instance, at the beginning of an initial session, the system responds to the command, "Where am I?" with the most orientation information, relating the user's current position to multiple major landmarks, such as the main gate, the station boundaries, etc. As the user progresses on the route, however, the position is described only in relation to the destination and a few landmarks in the nearby vicinity, e.g., "You are at the corner of Hudson and St. Lawrence. Hangar Number 3 is on your right. Transportation is on your left. You are approximately .8 miles southwest of your destination, the Geotechnical Laboratory."

Symmetry of Commands

Modifications to the voice input commands during the usability tests emphasized the importance of maintaining symmetry between input and output commands. For instance, when the command "Go Forward" was changed to "Continue Forward", the prompt for this command was changed from "To go forward one segment in your route, say 'go forward'." to "To continue forward one segment in your

route, say 'continue forward'." Other commands were modified similarly to improve the clarity of presentation.

Interruptibility

After reviewing several public domain synthesizers, a commercially available software synthesizer, TruVoice, produced by Centigram, Incorporated was chosen (Centigram 1996). It provided an acceptable vocal quality, but equally important, it allowed program control of the volume and speaking rate. Unfortunately, incompatibilities between its method of accessing the Sun audio device and the method used by other software components rendered the implementation of interruptibility beyond the scope of the current version of the program.

The lack of interruptibility presented the most significant limitation of the Speech Output Module. Compensating for this limitation required more stringent attempts to eschew auditory overload Again, responses were reviewed to ensure they were as brief as possible without leaving out necessary information. Likewise, lists were avoided or shortened wherever possible. For example, the help command "What can I ask" first attempts to give contextual help. If this does not suffice, it attempts to step the user through an incremental menu rather than giving all possible commands or queries. To illustrate, assume the user issues the command "What can I ask?" and has just heard the instructions for continuing on the next segment of the route, but has not continued forward. Three scenarios are possible: a) the user did not understand the directions and cannot remember how to get the program to repeat them, b) the user

cannot remember commands for navigating and traveling along the route, or c) the user cannot remember what queries can be made about the road or route Since the user has just heard the instructions, but has not continued forward, the help program assumes that either a) or b) must be true and thus prompts, "To hear the last instructions again, say 'repeat instructions'.", "To hear the commands for navigation and travel, say 'navigate'." If the user again says, "What can I ask?," incrementally more help is provided, e.g., the command for getting help regarding queries about the road, "road queries," is given.

Multimodal Interface

Design

This component of the prototype handles display of the visual and tactile map and coordinates its integration with the displayless interface. Although details in the design of the visual and tactile map differed, design of the underlying interface for both adhered to the same guidelines, including most importantly, simplicity, completeness, and immediacy.

Graphical Interface

The first critical decision in the design of the graphical interface resulted in foregoing the use of a Computer-Aided Design and Drafting (CADD)-GIS drawing (produced from the original database for scientists and engineers) for the visual map, using instead a visual map, designed by a graphic artist for visitors at the WES Visitor's Center. The map is shown in Appendix C. Since the latter was developed as a tool for

the general public, it met many of the design guidelines, offering a view of the station that was complete, yet simplified enough that an overview could be quickly grasped. Determining the areas of the map to be designated as "hot spots" for touch input presented the next design consideration. A desire to maintain simplicity motivated the decision to initially designate only laboratories as selectable, with the possibility of adding other areas after the usability tests.

Interactive Audio. The remaining design issues included determining the type of speech used in the output, synthetic or recorded, as well as the nature and amount of information to be presented. As Schmandt (1994) discusses, given a choice, listeners universally prefer the sound of a recorded human voice to a synthetic voice for numerous reasons, all related to its more natural and intelligible quality. Synthetic voices, however, are typically used in applications where recording all possible combinations of output is not practical, such as in the "WES Auto Travel" program. Its graphical interface, however, due to the presence of the visual map, needed only to provide audio for the selectable areas. Thus, the reduction in the length and variety of verbiage required made it feasible to use a recorded voice. The final issue concerned the nature and amount of information to provide. Recall that the enhanced prototype retains all functionality of the displayless prototype with the addition of graphical-auditory or tactile-auditory interaction. Therefore, the content of verbiage associated with the selectable areas served mainly as confirmation of other visual, tactile, and auditory cues, giving only the name of the laboratory and a short description of its location, e.g., "You have selected the Geotechnical Laboratory. It is located on

Hudson Road just inside the northeast gate to the station, approximately 1.3 miles northeast of the main entrance to the station." Participants in the usability tests preferred this redundancy because it served to confirm the current location or other directional information provided by the routing program.

Tactile Interface

As stated, design of this interface adhered to the same criteria as that of the graphical interface, with the exception that maintaining simplicity was much more critical. Tactile maps simply cannot represent the same level of detail as visual maps in a manner that is useful to the reader. Therefore, to determine the appropriate level of detail to provide, guidelines set forth in (Barth 1983) were studied and followed for the design of the tactile map. This required balancing the inherently conflicting goals of providing completeness while avoiding clutter. After careful consideration, it was determined that only vehicular roads referred to in the routing program would be represented on the map. Most footpaths or exclusively pedestrian walkways were too small to be tactually discernible and thus would only contribute to clutter. (Vehicular roads were categorized, however, and represented tactually as boundary roads or within-station roads, each with its own distinct texture.) Likewise, only laboratories were represented as landmarks on the map. Representing all landmarks required an overwhelming level of detail, given the physical space available for the map. The same principle dictated the manner of textual (Braille) labeling. The map did not provide sufficient space for full textual labeling of all objects. Therefore, as suggested in (Barth

1983) streets were labeled with the first two letters of the street name, e.g., "HU" for "Hudson," laboratories with the two initials of the lab name, e.g., "GL" for "Geotechnical Laboratory". Additionally, all textual labels were oriented horizontally, left to right, rather than vertically for ease of reading. Finally, a tactile legend explained the icons used to represent roads and laboratories as well as all textual abbreviations. Again, the entries in the key were presented in a horizontal, left to right orientation for reading purposes.

Implementation

Graphical Interface

To provide touch access to the map, the 21" Sun monitor included in the base hardware for the prototype was retrofitted with a touchscreen. It was important that the touchscreen provide pressure-sensitive, rather than capacitance-sensitive touch capabilities. Since the same hardware and software would be used for the tactile interface, pressure-sensitivity was necessary for users with visual impairments. This allowed tactile scanning and reading of the map without activating the audio. Software written in the tcl programming language controls the display of the visual map as well as its associated audio. Although the displayless and enhanced version of the prototype appear as one integrated program to the user, they are controlled by two separate, but communicating system-level processes.

Tactile Interface

Since the tactile interface uses the underlying software provided by the graphical interface, construction of the tactile map constituted the most significant task for implementation. Budgetary constraints dictated manual construction of a base map, which then served as a model for producing a raised image on thermoform paper. Choice of materials for the icons in the base map followed a trial-and-error process to determine which would produce the most readable image, in terms of size and sharpness of definition. For example, large pipe cleaners represented boundary roads, while smaller, thinner kite string represented within-station roads. The final thermoform image was placed in a detachable screen overlay situated on the perimeters of the touchscreen.

Prosodic Labeling Method

Requirements

Several factors influenced the selection of the transcription system to label the prosodics of the speech data collected in the experiments. First, the research required a transcription system that covered all aspects of prosody to be measured in the data, including tonal aspects as well as pauses and other durational aspects. Second, reporting the results in a manner accessible for peer review required use of a system that was accepted as a standard in the speech research community. Third, for feasibility, it was important that the selected system not demand specialized expertise or extensive training to begin use. Finally, though not a critical requirement,

availability of the system as a software tool compatible with the development environment was preferred. The TOBI transcription system, developed by a group of speech researchers from a variety of backgrounds in academia and industry, including speech recognition, speech synthesis, and computational linguistics, met all the aforementioned criteria, i.e., reliability, learnability and coverage. In addition, it was publicly available in a format supported by Entropic WAVES. Since it met and exceeded the requirements of the research, TOBI was chosen to perform prosodic labeling of the experimental data.

Transcription System

Silverman et al. (1992a) discuss evaluations of TOBI which demonstrate its reliability as a standard, showing levels of inter-transcriber agreement on evaluation tasks as high as 95%. (Lower levels were also reported, but many of the sources for inter-labeler disagreement were corrected after the evaluations. Silverman et al. (1992b) discuss this in further detail. In addition to assessing reliability, Silverman et al. (1992b) demonstrated the ease of learning TOBI. Given a speech corpus of 72 utterances, two transcribers with no experience and approximately one day's training achieved agreement rates as high as 94%. For further details, see (Silverman et al. 1992b).

Of equal importance to reliability and ease of learning, TOBI offers a structure for labeling several categories of prosody. This structure contains four tiers of transcription, an orthographic tier, a tonal tier, a break index tier, and a miscellaneous

tier. Each tier consists of a set of labels for denoting prosodic events as well as times at which these events occurred. The tiers and associated labeling conventions are described below.

Tonal Tier

Two types of tonal events are labeled in this tier, phrasal tones and pitch accents. For both events, the symbols "H" and "L" denote basic tone levels that are high or low, respectively, in the speaker's local pitch range. Phrasal tones can be either intermediate boundary tones, denoted with the symbol "-," or full intonational boundary tones, denoted with the symbol "%". For example, a low intermediate phrase boundary tone is denoted by the symbol "L-," while a high intonational boundary tone by the symbol "H%," and a low intermediate tone followed by a high intonational tone by the symbol "L-H%". An intermediate phrase boundary must always precede any full intonational phrase boundary, i.e., a full intonational phrase boundary encompasses, by definition, an intermediate phrase boundary. However, the reverse is not true, i.e., an intermediate phrase boundary can exist without being followed immediately by an intonational phrase boundary. The following example demonstrates this and how it is represented in the TOBI tonal tier :

Where am I and how far is it to Hydraulics?

L- H-H%

Note that the symbol "L-" is not followed by an intonational boundary symbol; however, the "H%" is preceded by an intermediate boundary symbol, "H-".

Pitch accents mark each accented syllable in an utterance. Labeling of pitch accents in TOBI is based on the intonational phonology of Pierrehumbert and Hirschberg (1990) with some modifications to simplify it for teaching and for use in automatic speech recognition. To summarize, TOBI uses 5 types of pitch accents, symbolized by "H*," "L*," "L*+H," "L+H*," and "H+!H*". The symbol "H*" denotes a peak accent relatively high in the speaker's range; "L*," a low accent in the lowest part of the speaker's range. The symbol "L*+H" denotes a low accented tone immediately followed by a sharp rise to the higher part of the speaker's range. Conversely, the symbol "L+H*" denotes a high peak on the accented syllable immediately preceded by a sharp rise from the lower part of the speaker's range. The symbol "H+!H" denotes an intonational downstep onto the accented syllable preceded by a high tone that was neither a high phrasal tone ending nor a high pitch accent. Detailed description of the pitch accents defined in TOBI is given in (Beckman and Ayers 1997). Though not used in the analysis, pitch accents were labeled in the experimental speech data for possible later investigations.

Break Index Tier

The break index tier measures the amount of disjuncture between adjacent words. This tier captures information on pauses that is not necessarily captured in the tonal tier, since pauses may accompany tonal cues or occur in the absence of tonal cues. It uses "break indices" based on the seven-point scale defined by (Price et al. 1991), but collapses this scale to a set of 4 break indices, with 0 representing the least

disjuncture and 4 representing the most. More specifically, the break index value of 0 represents clitic groups or words that have been closely grouped phonetically, such as in the phrase "did you?", spoken as "didjyou?". A break index value of 1 represents the level of disjuncture between two prosodic words, while a value of 2 represents a disjuncture or pause that is not accompanied by phrase tonal labeling. A break index value of 3 represents an intermediate phrase boundary, and value of 4 represents a full intonational phrase boundary.

Disfluencies are also represented in this tier. The symbol 'p' placed immediately after the break index denotes an audible hesitation and can only be used with break indices 1, 2, or 3. The symbol '1p' denotes a cutoff before a repair or restart. The symbol '2p' denotes a hesitation pause where no phrase accent is marked in the tonal tier, and '3p' denotes a hesitation pause where a phrase accent is marked in the tonal tier. All categories of break indices were used in both the labeling and analysis of the experimental data.

Orthographic Tier

This tier contains only a direct transcription of each word in the utterance with time alignments, but no prosodic information. This can be interfaced to dictionary entries that provide the typical lexical stress for a particular word since this is not included in TOBI.

Miscellaneous Tier

This tier serves as a 'comment' tier for labeling any events that cannot be categorized as belonging in the orthographic, break index, or tonal tier. This includes events such as coughing, laughing, or audible breaths.

CHAPTER V

EXPERIMENTAL RESULTS

Testing Conditions

Testing was conducted over a period of approximately three months at a variety of locations, including one major university, three rehabilitation agencies for the blind, and one sheltered industry for the blind. Multiple test sites were needed to elicit a sufficient sample size, particularly for subjects with visual impairments. The desire for diversity among the sample populations as well as practical considerations such as physical proximity, cost, and availability of subjects influenced the selection of test sites.

A total of 86 subjects participated in the experiments, including 30 sighted subjects, 28 subjects with adventitious vision loss and 28 subjects with congenital vision loss. (Subjects with vision loss indicated the time of onset of vision loss in a questionnaire, included in Appendix B.) Despite extensive usability testing of the prototype and careful subject selection, it was anticipated that some users might be unable to adapt to the interface within the time allotted for the experiment. This proved true for eight subjects. Some subjects were simply unable to adapt to the fundamental concepts and nature of the experiment even after the initial instructions

and warm-up session. These subjects either requested to end the experiment prematurely or required high levels of instructional assistance to complete the tasks. Thus, data from their sessions could not be used in the analyses. Other subjects exhibited certain vocal characteristics which created problems for the recognition component. More specifically, these characteristics precipitated unacceptable error rates by the recognizer. Therefore, data from sessions of these subjects could not be used since any prosodic features associated with cognitive stress may have been affected by the subject's frustration with recognizer errors. As an example, long-term smokers tended to breathe more audibly throughout an utterance, which was problematic for the recognizer, significantly increasing the error rate. Similar problems were presented by subjects with excessively soft voices or those lacking sufficient breath support. From all of the subjects who participated, data from sessions of 78 subjects were used in the analyses, including 27 sighted subjects, 25 subjects with adventitious vision loss, and 26 subjects with congenital vision loss.

The following sections detail results of the analyses. Overall session analyses are presented first, followed by task-level analyses as well as interpretation and discussion of the results. Results of additional analyses of the data for sighted subjects regarding cognitive preferences are then presented and interpreted. The chapter concludes with discussion of the relevance of the results for prosodic pattern detection algorithms.

Analyses of Data from Displayless Sessions versus Multimodal Sessions

It should be noted that the average number of utterances spoken in each session by subjects in all categories was approximately 60. In addition, the average number of words per utterance spoken in each session by subjects in all categories was approximately 10.

Subjects with Congenital Vision Loss

The results of analyses comparing all data from displayless sessions versus all data from the multimodal sessions for subjects with congenital vision loss are shown in Tables 1-8. To summarize the results, prosodic variables concerning pauses, intonational boundaries, and duration of utterances spoken by subjects in displayless sessions versus multimodal sessions differed at the significance level of alpha <= 0.05. Also, the number of recognition errors made by the system on utterances spoken by subjects differed significantly, alpha <= 0.05, in displayless sessions versus multimodal sessions. Prosodic variables concerning maximum and minimum F0 values, breaths, and disfluencies did not differ significantly between sessions. For all variables, a positive value in the table indicates the value for that variable was greater during the displayless session than the multimodal session; a negative value indicates the value for that variable was smaller during displayless sessions than multimodal sessions.

Pauses

The number of pauses in the "2p" category (occurring at a location other than a phrase boundary) as well as the "3p" category (occurring at a phrase boundary) was

significantly greater for utterances spoken when subjects used the displayless interface than the multimodal interface. These results are shown in Table 1.

Table 1. Matched-pair T Test Results for Number of Pauses per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	0.1538462	0.1976190	0.7784890	0.4436
2p	1.3461538	0.3840765	3.5049103	0.0017 ***
3p	2.0769231	0.8751669	2.3731736	0.0256 **

'**' Indicates mean difference was significant at alpha <= 0.05.
'***' Indicates mean difference was significant at alpha <= 0.01.

In addition, the average length of "2p" pauses of utterances spoken during displayless sessions was greater than during multimodal sessions at a significance level of alpha = 0.0561. These results are shown in Table 2.

Table 2. Matched-pair T Test Results for Average Pause Length per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	0.1141830	0.0862167	1.3243528	0.1974
2p	0.0856503	0.0427546	2.0032996	0.0561 *
3p	0.0110413	0.8941370	0.1234852	0.9027

'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.

Intonational Boundary Tones

The number of low full intonational boundary tones, denoted "L%", occurring in utterances spoken by subjects during displayless sessions was greater than during multimodal sessions, at a significance level of alpha = 0.0001. No other intonational boundary variables differed significantly between sessions. The results are shown below in Table 3.

Table 3. Matched-pair T Test Results for Number of Occurrences of Boundary Tones per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	0.1538420	0.7225858	0.2129106	0.8331
L%	16.8561538	3.4552971	4.8613880	0.0001 ***
H-	0.9230769	0.6322684	1.4599448	0.1568
H%	-0.3846154	1.9580814	-0.1964246	0.8459

'***' Indicates mean difference was significant at alpha <= 0.01.

Durational Features

Speaking rates were significantly slower for utterances spoken during displayless sessions than multimodal sessions. No other durational features differed significantly between the sessions. However, the increase in preboundary lengthening during displayless sessions, though not significant, alpha = 0.0722, was notable. The results for durational features are shown in Table 4.

Table 4. Matched-pair T Test Results for Durational Features per Utterance Spoken in Displayless vs.Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.0644100	0.0400754	1.6072436	0.1206
Speaking Rate **(Words/sec)**	-0.0982297	0.0437978	-2.2428018	0.0340 **
Preboundary Lengthening **(Syllables/sec)**	0.0559436	0.0297989	1.8773682	0.0722

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'**' Indicates mean difference was significant at alpha <= 0.05.

No other prosodic variables differed significantly between sessions in overall comparisons for subjects with congenital vision loss. Results for variables which did not differ significantly, including maximum and minimum F0 values, breaths, and disfluencies, are given in Tables 5-7. It is notable that maximum F0 differences varied widely; this is reflected in the standard error which is relatively high, approximately 24.95, with respect to the mean, approximately 2.45. Such a high variability contributed to the lack of significant differences for this feature.

Table 5. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	2.4541937	24.9506252	0.0983620	0.9224
Minimum	-1.4815462	1.64790470	-0.8990485	0.3772

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

Table 6. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions for Subjects with Congenital Vision Loss

N=26				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.4814815	0.5017643	0.9595770	0.3461
Non-boundary	0.1333330	0.4256622	1.0000000	0.3343

Table 7. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.3076923	0.2403154	1.2803688	0.2122

Recognition Errors

The number of substitution errors made by the system was significantly greater on utterances spoken by subjects during displayless sessions than during multimodal sessions. The number of rejection errors made by the system was also greater during displayless sessions than during multimodal sessions, at alpha = 0.0560. However, the number of insertion errors was fewer during displayless sessions than multimodal sessions, at alpha = 0.0570. Results for recognition errors are shown in Table 8.

Table 8. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Congenital Vision Loss

N=26				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	2.1153846	0.8214938	2.575043	0.0163 **
Insertion	-0.2307692	0.1151279	-2.0044593	0.0560 *
Rejection	0.5769231	0.2891530	1.9952172	0.0570 *

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.
'**' Indicates mean difference was significant at alpha <= 0.05.

Subjects with Adventitious Vision Loss

The results of analyses comparing all data from displayless sessions versus all data from multimodal sesions for subjects with adventitious vision loss are shown in Tables 9-16. To summarize, prosodic variables related to pauses, F0 features, and intonational boundary tones occurring in utterances spoken by subjects differed significantly in displayless sessions versus multimodal sessions. Recognition errors made by the system on utterances spoken by subjects also differed significantly in displayless versus multimodal sessions. Prosodic variables pertaining to breaths, disfluencies, and durational features did not differ significantly between sessions. For all variables, a positive value in the table indicates the value for that variable was greater during the displayless session than the multimodal session; a negative value

indicates the value for that variable was smaller during displayless sessions than multimodal sessions.

Pauses

The number of "2p" pauses in utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions at alpha = 0.0089. Also, the average length of "2p" pauses in utterances spoken by subjects was significantly longer during displayless sessions than multimodal sessions. Results for these variables are shown in Tables 9-10.

Table 9. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	0.04	0.2343786	0.1706640	0.8659
2p	1.00	0.3511850	2.8474740	0.0089 ***
3p	-0.64	0.8960655	-0.7142335	0.4820

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

Table 10. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	-0.0053588	0.0586496	-0.0913691	0.9280
2p	0.1745476	0.0786724	2.2186634	0.0326 **
3p	-0.1642300	0.1478045	-1.1111300	0.2775

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'**' Indicates mean difference was significant at alpha <= 0.05.

F0 and Intonational Features

Both minimum and maximum F0 values were significantly higher for utterances spoken by subjects during displayless sessions than during multimodal sessions with the increase in maximum F0 at a significance level of alpha = 0.0002. Results for F0 values are shown in Table 11. The number of full intonational boundary tones, both "L%" and "H%", was significantly greater for utterances spoken during displayless sessions than multimodal sessions. Results for "L%" boundary tones are at a significance level of alpha = 0.0006. Results for boundary tones are shown in Table 12.

Table 11. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	120.1876808	27.8282558	4.3189082	0.0002 ***
Minimum	14.70700000	7.07979479	2.0721300	0.0492 **

'**' Indicates mean difference was significant at alpha <= 0.05.
'***' Indicates mean difference was significant at alpha <= 0.01.

Table 12. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	-1.44	0.8776484	-1.6407482	0.1139
L%	15.08	3.8152850	3.95252250	0.0006 ***
H-	0.4	0.5354126	0.7470840	0.4623
H%	3.6	1.4364308	2.5062120	0.0194 **

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'**' Indicates mean difference was significant at alpha <= 0.05.
'***' Indicates mean difference was significant at alpha <= 0.01.

No other prosodic variables, including breaths, disfluencies, and durational features, differed significantly between for subjects with adventitious vision loss. Results for these variables are shown in Tables 13-15.

Table 13. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.36	0.2638181	1.3645765	0.1850
Non-boundary	1.52	1.6694510	0.9104789	0.3716

Table 14. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.3076923	0.2403154	1.2803688	0.2122

Table 15. Matched-pair T Test Results for Durational Features of Utterances Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.3312860	0.0318696	1.0395037	0.3089
Speaking Rate **(Words/sec)**	-0.0430569	0.0565359	-0.7615858	0.4537
Preboundary Lengthening **(Syllables/sec)**	0.0197710	0.0333722	0.5924391	0.5591

Recognition Errors

The number of substitution errors made by the system on utterances spoken by subjects was significantly higher during displayless sessions than during multimodal sessions at a significance level of alpha = 0.0010. No other categories of recognition errors made by the system on utterances spoken by subjects differed significantly between sessions. Results for recognition errors are shown in Table 16.

Table 16. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Subjects with Adventitious Vision Loss

N=25				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	2.96	0.8009994	3.6953837	0.0010 ***
Insertion	0.20	0.2336068	0.8944272	0.3800
Rejection	0.52	0.4550458	1.1427422	0.2644

'***' Indicates mean difference was significant at alpha <= 0.01.

Sighted Subjects

The results of analyses comparing all data from displayless sessions versus all data from multimodal sessions for sighted subjects are shown in Tables 17-25. To summarize the results, prosodic variables related to pauses, F0, intonational boundary tones and durational features of utterances spoken by subjects differed significantly in displayless versus multimodal sessions. No other prosodic variables, including breaths and disfluencies, differed significantly between sessions. However, recognition errors made by the system on utterances spoken by subjects differed significantly between sessions. For all variables, a positive value in the table indicates the value for that variable was greater during the displayless session than the multimodal session; a negative value indicates the value for that variable was smaller during the displayless session than the multimodal session.

Pauses

The number of "2p" pauses occurring in utterances spoken by subjects was greater during displayless sessions than during multimodal sessions at a significance level of alpha = 0.0001. The average length of "2p" pauses in utterances spoken by subjects was also significantly longer during displayless sessions than multimodal sessions at a significance level of alpha = 0.0057. Results for number and length of pauses are given in Tables 17-18.

Table 17. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.111111	0.2090089	-0.5316095	0.5595
2p	1.444444	0.3081668	4.6872167	0.0001 ***
3p	0.444444	0.7207414	0.6166490	0.5428

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

Table 18. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions

N=27				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	-0.1311974	0.0883212	-1.4854570	0.1494
2p	0.1483583	0.0491805	3.0166102	0.0057 ***
3p	0.0518285	0.0336261	1.5413161	0.1353

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

F0 and Intonational Features

Minimum F0 values occurring in utterances spoken by subjects were significantly lower during displayless sessions than during multimodal sessions. Also, the number of full intonational boundary tones occurring in utterances spoken by subjects, both high and low, "H%" and "L%", was significantly greater during displayless sessions than multimodal sessions with the increase in the "L%" category at alpha = 0.0007. Results for F0 and boundary tones are shown in Tables 19-20. Similar to the results for subjects with congenital vision loss, it is notable that maximum F0 values varied widely, which is reflected in the relatively high value for the standard error, approximately 19.9, with respect to the mean, approximately 5.4.

Table 19. Matched-pair T Test Results for Maximum and Minimum F0 Values Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	5.35540489	19.9130153	0.2689401	0.7901
Minimum	-5.55875940	1.7604692	-3.1575442	0.0040 ***

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

Table 20. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	-0.4074074	0.5925926	-0.6875000	0.4979
L%	13.7407407	3.5688557	3.8501811	0.0007 ***
H-	1.2962963	0.9693520	1.3372813	0.1927
H%	4.518585	2.2823305	1.9797828	0.0584 *

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.

Durational Features

The average duration of utterances spoken by subjects was significantly longer during displayless sessions than multimodal sessions at alpha = 0.0092. No other aspects of duration, including speaking rate or preboundary lengthening, differed significantly between sessions. Results for durational features are shown in Table 21.

Table 21. Matched-pair T Test Results for Durational Features of Utterances Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration (Sec)	0.1778601	0.0632127	2.8136774	0.0092 ***
Speaking Rate (Words/sec)	-0.0001486	0.0409151	-0.0036326	0.9971
Preboundary Lengthening (Syllables/sec)	0.0201271	0.0285863	0.7040815	0.4876

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'***' Indicates mean difference was significant at alpha <= 0.01.

No other prosodic variables , including breaths and disfluencies, differed significantly between sessions for sighted subjects. Results for breaths and disfluencies are shown in Tables 22-23.

Table 22. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.4814815	0.5017643	0.9595770	0.3461
Non-boundary	1.4074074	1.2139494	1.1593625	0.2568

Table 23. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.374041	0.1057590	0.7009042	0.4899

Recognition Errors

The number of substitution errors made by the system on utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions at a significance level of alpha = 0.0004. No other categories of errors differed significantly between sessions for sighted subjects. Results for recognition errors are shown in Table 24.

Table 24. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken in Displayless vs. Multimodal Sessions by Sighted Subjects

N=27				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	3.4444000	0.8461970	4.0704994	0.0004 ***
Insertion	0.3703704	0.2335661	1.5857196	0.1249
Rejection	0.0740741	0.4131938	0.1792720	0.8591

'***' Indicates mean difference was significant at alpha <= 0.01.

Summary

Certain trends can be observed in the overall session data for all categories of subjects. The number of "2p" or hesitation pauses, i.e., those which did not occur at a phrase boundary, occurring in utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions, at a significance level of alpha <= 0.01, for all populations. The average length of this type pause was also significantly greater during displayless sessions than multimodal sessions for sighted subjects as well as those with adventitious vision loss. It was also greater during displayless sessions than multilmodal sessions at a significance level of 0.05 <= alpha <= 0.06 for subjects with congenital vision loss. In addition, the number of "L%" boundary tones was significantly greater during displayless sessions, at a significance level of alpha <= 0.01, for all three populations. Finally, the number of substitution errors made by the system

on utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for all three populations as well. Beyond these basic trends, results differ for each of the three categories. Table 25 summarizes the prosodic variables which differ significantly for each subject category. The alpha value for each variable which differed is shown. A positive value indicates the variable was significantly larger during the displayless session, while a negative value indicates it was significantly smaller during the displayless session versus the multimodal session. Values of variables which differed at a significance level of 0.05 < = alpha <= 0.06 are marked with a single asterisk, '*'. Values of variables which differed at a significance level of alpha <= 0.05 are marked with a double asterisk, '**'. Values of variables which differed at a significance level of alpha <= 0.01 are marked with a triple asterisk, '***'.

Table 25. Summary of Significantly Differing Variables in Overall Sessions

SIGNIFICANT VARIABLES	Congenital	Adventitious	Sighted
Pauses			
Number 2p	0.0017 ***	0.0089 ***	0.0001 ***
Number 3p	0.0256 **		
Length 2p	0.0561 *	0.03260 **	0.0057 *
F0			
Maximum		0.0002 ***	
Minimum		0.0492 **	-0.0040 ***
Boundary Tones			
L-			
L%	0.0001***	0.0009 ***	0.0007 ***
H-			
H%		0.0526 *	0.0584 *
Durational Features			
Speaking Rate	-0.0340 **		
Duration			0.0092 ***
Errors			
Substitution	0.0163 **	0.0010 ***	0.0004 ***
Insertion	-0.0560 *		
Rejection	0.0570 *		

'-' Indicates value of variable was smaller during displayless session.
'***' Indicates mean difference was significant at alpha <= 0.01.
'**' Indicates mean difference was significant at alpha <= 0.05.
'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.

Several aspects of the results for subjects with congenital vision loss, at the overall level, differ from those of subjects with adventitious vision loss and those of sighted subjects. First, the number of "3p" pauses, i.e., those which occur at a phrase boundary, in utterances spoken by subjects is significantly greater during displayless sessions than multimodal sessions for this population. Second, F0 values do not change significantly between sessions. Third, only the number "L%" boundary tones is significantly greater during displayless sessions for this population, unlike the other two populations, for which the increase in the number of "H%" boundary tones during displayless sessions is significant also. Fourth, a greater number of durational features differ significantly between sessions for this population, and finally, all three categories of recognition errors differ significantly between sessions for this population. One interesting similarity between subjects with congenital vision loss and sighted subjects concerns the wide variability in the results for differences in maximum F0 values. These results, however, were produced from analyses of overall session comparisons. Any interpretation of the results as well as possible differences among populations must entail consideration of task-level analyses, presented below.

Task-level Analyses Comparing Data from Displayless Sessions Versus Multimodal Sessions

Two subjects completed all four tasks in each session. Five subjects completed three tasks in one or both sessions. All subjects, whose sessions were not eliminated from analyses for reasons described earlier, completed at least two tasks in one or both

sessions. Therefore, analyses of the task-level data were performed for the first two tasks only. To restate, the average number of utterances spoken in each session by subjects in all categories was approximately 60, with approximately one-third of the utterances in each session spoken for the first task and two-thirds of the utterances in each session spoken for the second task. In addition, the average number of words per utterance spoken in each session by subjects in all categories was approximately 10.

Subjects with Congenital Vision Loss

The results of the task-level analyses for subjects with congenital vision loss are given in Tables 26-41. Two subjects with congenital vision loss did not complete the second task in the second session. Therefore, all comparisons for the second task consider only 24 of the 26 subjects with congenital vision loss. Recall that prosodic variables pertaining to pauses, intonational boundary tones, and durational features differed significantly between sessions when considering all completed tasks for this population. Likewise, aspects of each of these variables differed at the task level as well.

Pauses

No significant differences between sessions for the number of pauses in utterances spoken by subjects were found for Task 1. However, for Task 2, the number of "2p" pauses was higher in utterances spoken by subjects during displayless sessions than during multimodal sessions, at a level of significance of alpha = 0.0024.

Also, the number of "3p" pauses in utterances spoken by subjects was significantly higher in displayless sessions than multimodal sessions for Task 2 only. Results comparing number of pauses for Task 1 and Task 2 are given in Tables 26-27.

Table 26. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.1153846	0.1280255	-0.9012627	0.3760
2p	0.4615385	0.2999014	1.5389675	0.1364
3p	0.3846154	0.4002957	0.9608281	0.3458

Table 27. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.2500000	0.1830696	1.3656006	0.1853
2p	0.9166667	0.2686296	3.4123821	0.0024 ***
3p	1.9167767	0.7915236	2.4214901	0.0237 **

'**' Indicates mean difference was significant at alpha <= 0.05.
'***' Indicates mean difference was significant at alpha <= 0.01.

The average length of "2p" pauses in utterances spoken by subjects was not significantly greater during displayless sessions than during multimodal sessions for Task 1 or Task 2, although in overall comparisons, this difference was significant at the level alpha = 0.0561. Results of comparisons of average pause length for Task 1 and Task 2 are given in Table 28-29.

Table 28. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	0.0124195	0.0348058	0.3568226	0.7242
2p	0.0597595	0.0385896	1.5485935	0.1340
3p	0.0207874	0.0813553	0.2555135	0.8004

Table 29. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	0.1176724	0.0998626	1.1783434	0.2507
2p	0.0456485	0.0422786	1.0797062	0.2915
3p	-0.0009420	0.0388478	-0.0242667	0.9808

Intonational Features

The number of "L%" intonational boundary tones in utterances spoken by subjects was significantly greater in displayless session than multimodal sessions for both Task 1 and Task 2. Although intermediate boundary tones did not differ significantly for overall session comparisons, the number of "L-" boundary tones in utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for Task 1 and the number of "H-" boundary tones was greater during displayless sessions than multimodal sessions at a significance level of alpha = 0.0549 for Task 1. Results are shown in Tables 30-31.

Table 30. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	-0.6153846	0.2606319	-2.3611253	0.0263 **
L%	3.4615385	1.5233877	2.2722636	0.0319 **
H-	0.5384615	0.2673561	2.0140237	0.0549 *
H%	0.6538462	0.8839462	0.7396900	0.4664

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.
'**' Indicates mean difference was significant at alpha <= 0.05.

Table 31. Matched-pair T Test Results for Number of Bounary Tones Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	0.458	0.6283725	0.7293975	0.4731
L%	9.500	3.2990996	2.8795736	0.0085 ***
H-	0.625	0.4920104	1.2702982	0.2167
H%	-1.250	1.4698048	-0.8504531	0.4038

'-' Indicates value was smaller during displayless session than during multimodal session.
'**' Indicates mean difference was significant at alpha <= 0.05.

Durational Features

The average speaking rate of utterances spoken by subjects was significantly slower during displayless sessions than multimodal sessions for Task 1 only. Similar to the overall comparisons, no other durational features differed significantly between sessions for Task 1 or Task 2. Results for durational features are shown in Tables 32-33.

Table 32. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Durational Features	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.0084919	0.0855300	0.992855	0.9217
Speaking Rate **(Words/sec)**	-0.1753698	0.0691238	-2.5370375	0.0178 **
Preboundary Lengthening **(Syllables/sec)**	0.0508505	0.0484787	1.0489255	0.3042

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 33. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Durational Features	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.1124590	0.6273010	1.7932213	0.0861
Speaking Rate **(Words/sec)**	0.0369308	0.0820148	0.4502946	0.6657
Preboundary Lengthening **(Syllables/sec)**	0.0430155	0.0358132	1.2011058	0.2419

In keeping with the overall results, no other prosodic variables, including F0 values, breaths, or disfluencies, differed significantly at the task level in displayless sessions versus multimodal sessions for subjects with congenital vision loss. Results for all variables which did not differ significantly are shown in Tables 34-39. Again, the wide variability in maximum F0 values was exhibited for both Task 1 and Task 2 and is reflected in the relatively high values for the standard error with respect to the mean.

Table 34. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
F0 Values	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	-3.1008865	20.8235954	-0.1489122	0.8828
Minimum	0.1539960	3.5585837	0.0432745	0.9658

Table 35. Matched-pair T Test Results for Maximum and Minimum F0 Values per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
F0 Values	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	12.2228743	24.7061889	0.4947292	0.6255
Minimum	-3.3757835	3.2540417	-1.0374125	0.3103

Table 36. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	-0.0384615	0.1412120	-0.2723674	0.7876
Non-boundary	0.2307692	0.2307692	1.0000000	0.3269

Table 37. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.416667	0.2686296	1.5510828	0.1345
Non-boundary	0.083333	0.8333440	1.0000000	0.3277

Table 38. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.1538462	0.2461538	0.625	0.5376

Table 39. Matched-pair T Test Results for Number of Disfluencies Occurring in Utterances Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.0384615	0.1034894	0.3716471	0.7133

Recognition Errors

The number of substitution errors made by the system on utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for Task 1 only, while the system made significantly fewer insertion errors during displayless sessions than multimodal sessions for Task 1 only. However, the number of rejection errors made by the system was significantly greater during displayless sessions than multimodal sessions for Task 2 only. Results are shown in Tables 40-41.

Table 40. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	1.3076923	0.6651067	1.9661393	0.0605 *
Insertion	-0.1538462	0.0721602	-2.1320072	0.0430 **
Rejection	-0.1153846	0.1782558	-0.6472978	0.5233

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.
'**' Indicates mean difference was significant at alpha <= 0.05.

Table 41. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=24				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	1.0000000	0.6454972	1.5491933	0.1350
Insertion	-0.0833333	0.0576303	-1.4459976	0.1617
Rejection	0.6666667	0.2801828	2.3793988	0.0260 **

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'**' Indicates mean difference was significant at alpha <= 0.05.

Subjects with Adventitious Vision Loss

The results of task-level analyses for subjects with adventitious vision loss are given in Tables 42-57. Five of the subjects with adventitious vision loss did not complete Task 2 in the second session. Therefore, analyses for Task 2 consider only 20 of the 25 subjects with adventitious vision loss. Recall that variables pertaining to pauses and intonation as well as system recognition errors differed significantly in overall session comparisons. These variables also differed significantly at the task level.

Pauses

The number of "2p" pauses occurring in utterances spoken by subjects during displayless sessions was significantly greater than during multimodal sessions for Task 2 only. Also, the average length of "2p" pauses was significantly longer during displayless sessions than multimodal sessions for Task 2 only. These results are shown in Tables 42-45.

Table 42. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	0.12	0.1854724	0.6469966	0.5238
2p	0.20	0.1632993	1.2247449	0.2326
3p	-0.60	0.5744563	-1.0444659	0.3067

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

Table 43. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.1052632	0.2007987	-0.5242224	0.6065
2p	1.0526316	0.3859649	2.7272727	0.0138 **
3p	-0.0526316	0.8001231	-0.0657794	0.9483

'-' Indicates value of variable was smaller during displayless session than during multimodal session.

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 44. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.0144299	0.0424330	-0.3400630	0.7368
2p	0.0378616	0.0518960	0.7295663	0.4727
3p	-0.0819818	0.1053219	-0.7783932	0.4439

Table 45. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	0.0119357	0.0730812	0.1633207	0.8721
2p	0.1798500	0.0755227	2.3814019	0.0285 **
3p	-0.1082212	0.1455289	-0.7436407	0.4667

'**' Indicates mean difference was significant at alpha <= 0.05.

Intonational Features

Maximum F0 values were significantly higher in utterances spoken during displayless sessions than multimodal sessions for Task 1 and Task 2. However, minimum F0 values were higher in utterances spoken during displayless sessions than multimodal sessions for Task 1 only. These results are shown in Tables 46-47.

Table 46. Matched-pair T Test Results for Minimum and Maximum F0 Values Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	67.0362432	27.051081	2.4781412	0.0206 **
Minimum	14.4970553	6.778727	2.1386101	0.0428 **

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 47. Matched-pair T Test Results for Minimum and Maximum F0 Values Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	71.4516353	23.9905982	2.9783182	0.0081 ***
Minimum	0.0902576	2.2167096	0.0407169	0.9680

'***' Indicates mean difference was significant at alpha <= 0.01.

The number of "L%" boundary tones in utterances spoken by subjects during displayless sessions was significantly greater than during multimodal sessions for both Task 1 and Task 2. For Task 1, the number of "L%" boundary tones differed at a significance level of alpha = 0.0009. The number of "H%" boundary tones was

significantly greater for utterances spoken by subjects in displayless sessions versus multimodal sessions for Task 1 only. The number of "H-" boundary tones, which did not differ significantly at the overall level, was significantly greater for utterances spoken by subjects during displayless sessions than multimodal sessions for Task 2 only. The results for boundary tones are shown in Tables 48-49.

Table 48. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	-0.345144	0.1795100	-1.9226974	0.0705
L%	11.360000	3.0121310	3.7714630	0.0009 ***
H-	-0.400000	0.3316625	-1.2060454	0.2396
H%	2.12	1.0397436	2.0389643	0.0526 **

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.
'***' Indicates mean difference was significant at alpha <= 0.01.
'**' Indicates mean difference was significant at alpha <= 0.05.

Table 49. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	0.1834775	0.4132003	0.4440401	0.6623
L%	6.8333333	2.6349213	2.5933728	0.0189 **
H-	1.1666667	0.5377878	2.1663812	0.0445 **
H%	2.1666667	1.7734407	1.2492257	0.2285

'**' Indicates mean difference was significant at alpha <= 0.05.

Durational Features

Although no durational features differed significantly in overall session comparisons, preboundary lengthening was significantly greater in utterances spoken by subjects during displayless sessions than multimodal sessions for Task 2 only. Results for durational features are shown in Tables 50-51.

Table 50. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration (Sec)	0.0749284	0.0577778	1.2968371	0.2070
Speaking Rate (Words/sec)	-0.0945175	0.0699186	-1.3512422	0.1892
Preboundary Lengthening (Syllables/sec)	0.0036016	0.0453867	0.0793545	0.9374

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

Table 51. Matched-pair T Test Results for Durational Features of Utterances Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Durational Features	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.0055254	0.0535248	0.1032315	0.9189
Speaking Rate **(Words/sec)**	-0.0476754	0.0664595	-0.7127361	0.4819
Preboundary Lengthening **(Syllables/sec)**	0.0698471	0.0317937	2.1968843	0.0406 **

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'**' Indicates mean difference was significant at alpha <= 0.05.

No other prosodic variables, including breaths and disfluencies, differed significantly between sessions at the task level for subjects with adventitious vision loss. Results for these variables are shown in Tables 52-55.

Table 52. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.24	0.2400000	1.0000000	0.3273
Non-boundary	1.04	1.3459569	0.7726845	0.4473

Table 53. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.166667	0.1666670	1.0000000	0.2644
Non-boundary	1.444444	1.8755779	0.7701330	0.4518

Table 54. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.320	0.1451436	0.826764	0.4165

Table 55. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	-0.0555556	0.1709684	-0.3249462	0.7992

Recognition Errors

The number of substitution errors made by the system on utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for both Task 1 and Task 2. In keeping with the overall results, no other categories of recognition errors made by the system differed significantly at the task level. Results are shown in Tables 56-57.

Table 56. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=25				
Error Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	1.80	0.7071068	2.5455844	0.0178 **
Insertion	-0.04	0.0909212	-0.4399413	0.6639
Rejection	0.60	0.4320494	1.3887301	0.1777

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 57. Matched-pair T Test Results for Number of Recognition Errors Made by the System per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Error Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	2.36	0.6579767	3.5867532	0.0015 ***
Insertion	0.32	0.18	1.7777778	0.0881
Rejection	0.16	0.1796292	0.8907235	0.3819

'***' Indicates mean difference was significant at alpha <= 0.01.

Sighted Subjects

The results of task-level analyses for sighted subjects are given in Tables 59-71. All sighted subjects completed at least two tasks in each session. Therefore, analyses for Task 2 include all 27 sighted subjects. Recall that prosodic variables pertaining to pauses, F0, intonational boundary tones, and durational features differed significantly in overall session comparisons. In addition, the number of recognition errors made by the system differed significantly at the overall level. Each of these variables differed at the task level as well.

Pauses

The number of "2p" pauses occurring in utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for both Task 1 and Task 2. The average length of "2p" pauses was also significantly greater in

utterances spoken by subjects during displayless sessions than multimodal sessions, but for Task 2 only. Results for pauses are shown in Tables 58-61.

Table 58. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	-0.111111	0.1343268	-0.8271702	0.4157
2p	0.370374	0.1523617	2.4308622	0.0223 **
3p	-0.407407	0.2628967	-1.5496800	0.1333

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 59. Matched-pair T Test Results for Number of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Pause Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
1p	0.0370370	0.1249670	0.2963745	0.7693
2p	0.8888889	0.2465334	3.6055513	0.0013 ***
3p	1.1481481	0.6953318	1.6512234	0.1107

'***' Indicates mean difference was significant at alpha <= 0.05.

Table 60. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	-0.04134760	0.0354692	-1.1657349	0.2543
2p	0.02962963	0.1755319	1.6879914	0.1034
3p	-0.01650750	0.0418678	-0.3942772	0.6966

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

Table 61. Matched-pair T Test Results for Average Length of Pauses Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Pause Type	**Mean Diff. (Sec)**	**Std. Error**	**T**	**Prob(T)**
1p	-0.0224223	0.0480714	-0.4664368	0.6448
2p	0.8148148	0.2389259	3.4103247	0.0021 ***
3p	0.0468873	0.0418273	1.1209738	0.2725

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'***' Indicates mean difference was significant at alpha <= 0.01.

Intonational Features

Minimum F0 values were significantly lower per Utterance Spoken by subjects during displayless sessions than multimodal sessions for both Task 1 and Task 2, at a significance level of alpha = 0.0057. Similar to the overall results, maximum F0 values did not differ significantly between sessions at the task level. In addition, the wide variability exhibited at the overall level was exhibited at the task level, as indicated by the relatively high value for standard error with respect to the mean. Results for minimum and maximum F0 values are shown in Tables 62-63.

Table 62. Matched-pair T Test Results for Minimum and Maximum F0 Values Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
F0 Values	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	-5.3693691	16.9277844	-0.3171927	0.7536
Minimum	-9.9597735	3.4167324	-2.914996	0.0072 ***

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'***' Indicates mean difference was significant at alpha <= 0.01.

Table 63. Matched-pair T Test Results for Minimum and Maximum F0 Values Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Maximum	4.4699651	25.2036614	0.1773538	0.8606
Minimum	-5.1345149	1.7036368	-3.0138553	0.0057 ***

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'***' Indicates mean difference was significant at alpha <= 0.01.

The number of "L%" boundary tones occurring in utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for both Task 1 and Task 2; however, the number of "H%" boundary tones was significantly greater for utterances spoken in displayless sessions than multimodal sessions for Task 2 only. These results are shown in Tables 64-65.

Table 64. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	-0.2962963	0.2122651	-1.3958788	0.1746
L%	7.0000000	2.8470007	2.4587279	0.0209 **
H-	-0.3333333	0.2445998	-1.3627703	0.1846
H%	0.0000000	0.8059607	0.0000000	1.0000

'-' Indicates value of variable was smaller during displayless sessions than during multimodal sessions.

'**' Indicates mean difference was significant at alpha <= 0.05.

Table 65. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Boundary Tone Type	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
L-	0.037	0.4637316	0.0798674	0.9370
L%	10.960	2.7967632	3.9198752	0.0006 ***
H-	0.625	0.4920104	1.2702982	0.2167
H%	3.330	1.5824897	2.1063855	0.0450

'***' Indicates mean difference was significant at alpha <= 0.01.

Durational Features

The average duration of utterances spoken by subjects during displayless sessions was significantly longer than during multimodal sessions for Task 2 only, but notable for Task 1, alpha = 0.0750. Similar to the overall comparisons, no other durational features differed significantly between sessions at the task level. Results are shown in Tables 66-67.

Table 66. Matched-pair T Test Results for Durational Features of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration **(Sec)**	0.1804655	0.0972931	1.8548645	0.0750
Speaking Rate **(Words/sec)**	-0.1268214	0.0933649	-1.3583417	0.1860
Preboundary Lengthening **(Syllables/sec)**	0.0340450	0.0648855	0.5246935	0.6042

Table 67. Matched-pair T Test Results for Durational Features of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Durational Feature	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Total Duration (Sec)	0.1762502	0.0574138	3.069822	0.0050 ***
Speaking Rate (Words/sec)	-0.0568371	0.0721680	0.7875662	0.4381
Preboundary Lengthening (Syllables/sec)	0.0147753	0.0437359	0.3378301	0.7382

'***' Indicates mean difference was significant at alpha <= 0.01.

No other prosodic variables, including breaths and disfluencies, differed significantly between sessions at the task level for sighted subjects. These results are shown in Tables 68-71.

Table 68. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Breath Location	**Mean Difference**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.00	0.0533761	0.00	1.0000
Non-boundary	1.48	1.0368844	1.43	0.1650

Table 69. Matched-pair T Test Results for Number of Breaths Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Breath Location	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Boundary	0.0740741	0.2686296	1.5510828	0.1345
Non-boundary	0.3703704	0.2781100	1.3317407	0.1945

Table 70. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.3076923	0.2403154	1.2803688	0.2122

Table 71. Matched-pair T Test Results for Number of Disfluencies Occurring per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Disfluencies	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
	0.0740741	0.0913015	0.8113124	0.4246

Recognition Errors

The number of substitution errors made by the system on utterances spoken by subjects was significantly greater during displayless sessions than multimodal sessions for Task 2 at a significance level of alpha = 0.0072, but was not significantly greater in displayless sessions than multimodal sessions for Task 1. Results are shown in Tables 72-73.

Table 72. Matched-pair T Test Results for Number of Recognition Errors Made By the System per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 1

N=27				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	1.04	0.6645209	1.5650161	0.1307
Insertion	0.32	0.2628054	1.2176311	0.2352
Rejection	0.00	0.1632993	0.0000000	1.0000

Table 73. Matched-pair T Test Results for Number of Recognition Errors Made By the System per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions for Task 2

N=27				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	2.12	0.7218495	2.9369	0.0072 ***
Insertion	0.00	0.0816497	0.0000	1.0000
Rejection	0.05	0.3372437	1.4233	0.1675

'***' Indicates mean difference was significant at alpha <= 0.01.

Discussion of Task-level Results

To review, task-level analyses were performed to determine the extent to which the spatial complexity of the tasks may have impacted the user's cognitive load and hence, prosodics. Such an examination assumes that significant differences at the overall level which are found significant for Task 2 support the research hypothesis more strongly than those found significant for Task 1 only. Further, those differences which only become significant or increase in significance for Task 2 provide strongest support for the research hypothesis.

Variables which differed significantly on Task 2 for all populations include the number of "2p" pauses and the number of "L%" boundary tones, all of which were significantly greater for utterances spoken during displayless sessions than multimodal sessions. Many of the trends which distinguished each population at the overall level

surfaced at the task level as well. However, not all remained significant for Task 2. The distinguishing features for each population are described below.

Subjects with Congenital Vision Loss

For subjects with congenital vision loss, an increase in the average length of "2p" pauses in utterances spoken during displayless sessions versus multimodal sessions was not found significant for Task 1 or Task 2, although it was found significant in overall session comparisons. The number of "3p" pauses was significantly greater in utterances spoken during displayless sessions than multimodal sessions for Task 2 only.

Durational features, including speaking rate and duration of utterance did not differ significantly for Task 2. The decrease in speaking rate of subjects for Task 1 only, and not Task 2, could have been caused by the behavior of subjects who were unable to complete Task 2 in the multimodal session. Table 74 shows the results of removing those subjects from the analyses for Task 1.

Table 74. Matched-pair T Test Results for Speaking Rate per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=24				
Durational Features	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Speaking Rate (Words/sec)	-0.1642294	0.0731551	-2.2449473	0.0347 **

'**' Indicates mean difference was significant at alpha <= 0.05.

Removing the subjects who did not complete Task 2 in the multimodal session affected the significance level of the decrease in speaking rate during the displayless sessions, shifting it from alpha = 0.0178 to alpha = 0.037. However, alpha remained less than 0.05, providing no statistical evidence that the behavior of the subjects unable to complete Task 2 in the multimodal session caused the decrease in speaking rate on Task 1 during the displayless sessions.

Regarding recognition errors made by the system, only rejection errors, not substitution or insertion errors, were significantly greater on utterances spoken for Task 2 during the displayless sessions, although all types of recognition errors made by the system differed significantly at the overall level. Also, both substitution and insertion errors were significantly greater on utterances spoken for Task 1 during the displayless sessions.

Additional analyses were conducted for Task 1, removing from consideration those subjects who did not complete Task 2 in the multimodal sessions. These analyses were performed in order to examine the potential impact of these subjects on the increase in errors made by the system for Task 1 only, and not Task 2. Table 75 shows the results of these analyses.

Table 75. Matched-pair T Test Results for Substitution and Insertion Errors Made by the System per Utterance Spoken by Subjects with Congenital Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=26				
Error Category	**Mean Diff.**	**Std. Error**	**T**	**Prob(T)**
Substitution	1.5416667	0.6941243	2.2210238	0.0365 **
Insertion	-0.125000	0.0689597	-1.8126539	0.0619**

'-' Indicates value of variable was smaller during displayless session than during multimodal session.
'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.
'**' Indicates mean difference was significant at alpha <= 0.05.

Removing subjects who did not complete Task 2 in the multimodal sessions from consideration raised the significance level of the increase in substitution errors during displayless sessions for Task 1 from alpha = 0.0605 to alpha = 0.03655. This offers statistical evidence contrary to an argument that the behavior of subjects unable to complete Task 2 in the multimodal sessions caused the increase in substitution errors for Task 1 in the displayless sessions. The significance level of the increase in insertion errors during displayless sessions, however, shifted from alpha = 0.0430 to alpha = 0.0619. Although this increased alpha to a value greater than 0.05, it remained notable at 0.05 <= alpha <= 0.06, which does not strongly support an argument that the subjects unable to complete Task 2 in the multimodal session caused the increase in insertion errors for Task 1 in the displayless sessions.

Subjects with Adventitious Vision Loss

For subjects with adventitious vision loss, the maximum F0 value was significantly higher in utterances spoken for Task 2 during displayless sessions than multimodal sessions; however, the minimum F0 value was not significantly higher in utterances spoken during displayless sessions for Task 2, but was significantly higher for Task 1.

Further analyses were conducted for Task 1, removing from consideration those subjects who did not complete Task 2 in the multimodal sessions. The additional analyses were conducted in order to examine the impact of these subjects on the increase in minimum F0 values during displayless sessions for Task 1 only, and not Task 2. Table 76 shows the results of these analyses. The alpha value shifted from 0.0428 to 0.0623, creating a reduction in the significance level for the increase in minimum F0 values during displayless sessions for Task 1. Again, however, alpha remained at a notable level, 0.05 <= alpha <= 0.06, which does not strongly indicate that the behavior of the users unable to complete Task 2 in the multimodal sessions caused the increase in minimum F0 values for Task 1 of the displayless sessions.

Table 76. Matched-pair T Test Results for Minimum F0 Values Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 1

N=20				
F0 Value	**Mean Diff. (Hz)**	**Std. Error**	**T**	**Prob(T)**
Minimum	13.0752451	6.9101064	1.8921916	0.0623*

'*' Indicates mean difference was significant at 0.05 <= alpha <= 0.06.

The number of "H%" boundary tones, which was significantly higher in utterances spoken during displayless sessions than multimodal sessions at the overall level, did not remain so for Task 2, although it was significant for Task 1. Additional analyses were performed, removing from consideration subjects who did not complete Task 2 in the multimodal sessions. These analyses were conducted in order to examine the effect of the behavior of these subjects on the increase in the number of "H%" boundary tones during displayless sessions for Task 1 only, and not Task 2. Table 77 shows the results of these analyses. The alpha value shifted from 0.0526 to 0.0783, which represents a reduction in the significance level. While this indicates some impact by the subjects who did not complete Task 2 in the multimodal session, it does not clearly show that these subjects caused the increase in the number of "H%" boundary tones during displayless sessions for Task 1.

Table 77. Matched-pair T Test Results for Number of Boundary Tones Occurring per Utterance Spoken by Subjects with Adventitious Vision Loss in Displayless vs. Multimodal Sessions for Task 2

N=20				
Boundary Tone Type	**Mean Difference**	**Std. Error**	**T**	**Prob(T)**
H%	1.875	1.0534837	1.7798092	0.0783

'**' Indicates mean difference was significant at alpha <= 0.05.

The number of "H-" boundary tones, not significantly greater for utterances spoken during displayless sessions than multimodal sessions at the overall level, was greater for Task 2 at the significance level, alpha = 0.0586. Interestingly, preboundary lengthening, a variable which did not differ significantly at the overall level, was significantly greater in utterances spoken for Task 2 during displayless sessions than multimodal sessions.

Sighted Subjects

For sighted subjects, the minimum F0 value was significantly lower in utterances spoken for Task 2 during the displayless sessions than multimodal sessions at the significance level, alpha <= 0.01. However, the maximum F0 value did not differ significantly and showed a high level of variation for both Task 1 and Task 2. The number of "H%" boundary tones was significantly greater in utterances spoken for Task 2 during displayless sessions than multimodal sessions. Finally, the total duration of utterances spoken for Task 2 was significantly greater during displayless sessions

than multimodal sessions; also, the number of substitution errors made by the system was significantly greater on utterances spoken by subjects for Task 2 during displayless sessions than multimodal sessions, at a significance level of alpha <= 0.01.

Table 78 summarizes the task-level results for variables which differed significantly between sessions at the overall level for each population. The alpha value for each variable which differed is shown. A positive value indicates the variable was significantly larger during the displayless session, while a negative value indicates it was significantly smaller during the displayless session versus the multimodal session. Variables which differed at a significance level of 0.05 <= alpha <= 0.06 are marked with a single asterisk, '*'. Variables which differed at a significance level of alpha <= 0.05 are marked with a double asterisk, '**'. Variables which differed at a significance level of alpha <= 0.01 are marked with a triple asterisk, '***'.

Table 78. Summary of Task-level Results for Variables Differing Significantly in Overall Sessions

Variables	Congen. Task 1	Congen. Task 2	Advent. Task 1	Advent. Task 2	Sighted Task1	Sighted Task 2
Pauses						
Number 2p		0.0024***		0.0138**	0.0233**	0.0013**
Number 3p		0.0237**				
Length 2p				0.0285**		0.0021**
F0						
Maximum			0.0206**	0.0081***		
Minimum			0.0428**		-0.0061***	-0.0057***
Boundary Tones						
L-						
L%	0.0319**	0.0085**	0.0009**	0.0189**	0.0209**	0.0006**
H-				0.0586 *		
H%			0.0526 *			0.0450**
Durational Features						
Speaking Rate	-0.0178**					
Duration						0.0050***
Preb. Length.				0.0406**		
Recog. Errors						
Substitution	0.0605*		0.0178**	0.0015***		0.0072***
Insertion	-0.0430**					
Rejection		0.0260**				

'-' Indicates value of variable was smaller during displayless session.
'***' Indicates difference was significant at alpha <= 0.01.
'**' Indicates difference was significant at alpha <= 0.05.
'*' Indicates difference was significant at 0.05 <= alpha <= 0.06.

Interpretation of Results

There is strong evidence in the data, at the task level, that hesitation pauses at locations other than phrase boundaries, i.e., "2p" pauses, are increased in the displayless condition versus the multimodal condition for all populations. This finding is important for several reasons. First, since the average number of utterances per session as well as words per utterance remained the same in both sessions, this increase cannot be attributed to users simply speaking less, i.e., fewer utterances or fewer words per utterance, in the multimodal sessions than the displayless sessions. Therefore, these pauses likely indicate an increase in cognitive effort by the user when employing the displayless interface to perform a navigational task. This additional cognitive load, if not offset by other factors, can lead to a decrease in user satisfaction with the interface. More specifically, this type of pause appears to have increased the error rate of the recognizer, as evidenced in the increase in substitution errors, which also negatively impacts user satisfaction.

Human Factors Issues

The dissimilarities in the results for the tonal data as well as other aspects of pauses for subjects with congenital vision loss, however, suggest a more complex relationship between prosodic features and recognition error rates. For this population only, substitution errors were not significantly increased during displayless sessions for Task 2. Other trends which distinguish this population include a lack of significant differences in F0 values between sessions, fewer differences in intonational boundaries,

and fewer differences in pauses, i.e., the length of hesitation pauses was not significantly greater during displayless sessions. Also, there were significantly more "3p" pauses in the displayless sessions. The correlation between these variations in prosodic features and the recognition error rate bears further exploration.

Recognition Errors and Prosodics

Examination of the results for the sighted and adventitious populations provides additional insight into the relationship between prosodic features and the substitution error rate. As noted, the results for the sighted population showed significant changes in intonational features between sessions as well as a significant increase in substitution errors during the displayless session. However, the data for the adventitious population exhibited the greatest number of differences between sessions in tonal data (both maximum and minimum F0 and all categories of intonational boundary tones) as well as the greatest increase in substitution errors during the displayless session. This suggests that the interaction between intonational features and hesitation pauses may have produced the greatest effect on the substitution error rate.

Rosenfeld et al. (1996) examined the correlation between disfluencies and recognition error rate and found that disfluencies did not significantly impact the word recognition error rate. Two characteristics of their work, however, make it less applicable to this research. First, the study examined disfluencies, not pauses exclusively. Second, their research examined an application called "Switchboard," a database of speech produced from telephone conversations between two speakers.

Spontaneous conversations of this nature tend to include many monosyllabic words and phrases, such as "yeah" and "huh?". In fact, monosyllabic words covered 75% of the Switchboard database. In addition, these monosyllabic utterances accounted for over 80% of the recognition errors, which tends to mask any errors produced by disfluencies. Users of the "WES Auto Travel" application, however, employ structured natural language queries to perform specific tasks. This reduces the possibilities for monosyllabic phrases in an utterance as well as the errors they introduce. The differences in applications and in the linguistic phenomena measured, i.e., pauses versus all disfluencies, increases the value of an additional study using the data from this research to explore the relationship between specific prosodic features and recognition error rate.

Prosodics and Cognitive Load

Again, the dissimilarities in prosodic variations among populations in this research are important because they serve to elucidate the relationship between prosodics and the recognition error rate. Of equal importance, however, the dissimilarities strongly suggest differences in the way in which each population adapts to a displayless navigational interface. In particular, assuming that significant changes in prosodics reflect additional cognitive load, subjects in each population experienced additional cognitive load, since each exhibited significant prosodic variations. Further, subjects in the sighted and adventitious populations, exhibiting the most prosodic changes during the displayless session, experienced greatest additional cognitive load in

the absence of a visual or tactile display. Likewise, subjects in the congenital population, exhibiting fewer prosodic variations during the displayless session, experienced less additional cognitive load in the absence of the visual or tactile map. It should be noted that, in general, these subjects tended to more quickly locate speech shortcuts for performing tasks than sighted subjects or subjects with adventitious vision loss. However, no formal analysis of this phenomenon was conducted. Further study of this issue and how it may relate to lack of visual memory is needed.

Cognitive Load

Perhaps more importantly, an investigation of methods for reducing the cognitive load for all users of displayless interfaces should be conducted. Observations of subjects from this study provide a basis for such an investigation. Subjects appeared to have the most difficulty simply maintaining a general sense of compass directions when using the displayless interface. Although the program provides explicit instructions regarding compass directions and whether to turn left, right, or continue, subjects were continually translating this information with respect to their current location, particularly when forming their own queries. (For many subjects, the translation process could be physically observed, either through verbalizing, e.g., "If north is to my left, I'll go south if I turn right.", gesturing, e.g., using fingers to trace position in the air, closing eyes to "visualize" the area for sighted subjects, or a combination of verbalizing and gesturing.)

Assuming that this translation effort increased the cognitive load of the user, two possible approaches for ameliorating this problem should be examined. First, as previously argued, in many situations a displayless interface is necessary because the use of any sort of visual or tactile map is not feasible. However, in certain situations, it may be possible to employ a visual or tactile map, greatly reduced in size, e.g., palm size or head-mount display size. Allowing the user to refer to such a map for general compass directions could reduce some of the burden created in translating strictly verbal directions. Although the maps used in the experiments for this research were kept as simple as possible, they contained much more detail than would be provided in such a map. The extent to which this map, in combination with speech, might reduce cognitive load, and hence affect speaker prosodics, should be examined.

Of course, in many situations, it is not feasible to provide any sort of map. For these applications, augmenting verbal directions with non-speech audio cues could potentially reduce cognitive load. In particular, using stereo localization cues to convey the direction of travel could potentially facilitate the translation process. For example, assuming "north" is located to the user's left, then a spoken direction to move northward would place audio cues in the user's left earphone. Loomis et al. (1994) describe research which implemented such techniques in a virtual acoustic display used in a personal guidance system for the visually impaired. Although they believed this technique would ultimately facilitate the user in constructing a mental map of an area, they encountered several practical problems in its implementation. Assuming these

problems are overcome, implementing such an interface with a speech recognition component would allow formal study of how localized sound affects cognitive load and speaker prosodics.

To restate, the issue of cognitive load is important because it directly impacts user satisfaction with the interface. For speech recognition interfaces, however, the issue of recognition errors is similarly important since it also directly impacts user satisfaction. Prosodics interlinks these issues, both in its reflection of cognitive load and its potential impact on recognition error rates. The relationship among these four variables, cognitive load, prosodics, recognition errors, and user satisfaction, is illustrated in Figure 5.

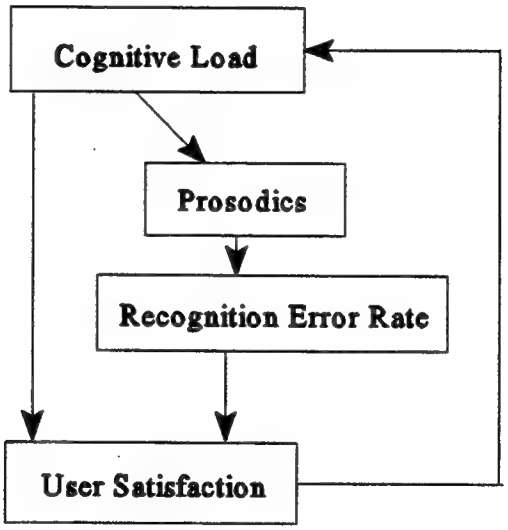


Figure 8. Relationship Among Cognitive Load, Prosodics, Error Rate, and User Satisfaction

In view of the role of recognition errors in this relationship, the significant increase in rejection errors for the congenital population on task two of the displayless session cannot be ignored. This increase is difficult to explain since these types of errors are typically caused by out-of-vocabulary utterances, which tend to lessen as the user adapts to the interface. Speech produced from the completion of additional tasks could provide better insight into this issue. However, as discussed, only two tasks could be analyzed since the sample size for those who completed more than two tasks was not sufficient for analysis. This leads to a review of some limitations of this study and more generally, the problems and costs associated with data collection.

Limitations of Study

The cost of data collection frequently presents unavoidable constraints in the research process. While this may be particularly true for research involving human subject testing, several issues specific to this study should be mentioned. First, computer users with visual impairments constitute a low incidence population. Consequently, greater resources are required to locate and recruit subjects for testing than those required for a higher incidence population. In this study alone, five agencies in three different states were visited to elicit a sample size sufficient for analysis. This clearly impacted the resources needed for planning, coordination, and travel to conduct the experiments.

Second, all subjects, regardless of visual capability, required more time than that allotted to complete all four tasks in each session. Allotting additional time to

complete the experiment could have increased the amount of time required from each subject by at least twofold or greater. As conducted, the experiment required approximately two hours per subject including time to perform the experiment and respond to the questionnaires. As the amount of time required from subjects to participate in an experiment expands, the importance and level of remuneration increases. This presents an additional budgetary constraint.

Third, time appears to have presented a limitation for some subjects regarding use of the tactile map. Two subjects with congenital vision loss and five subjects with adventitious vision loss could not complete the second task in the second session, which entailed use of the tactile map. Several observations support the hypothesis that lack of experience with tactile maps increased the time needed to complete the tasks for these subjects. First, subjects with adventitious vision loss comprised the largest number of subjects who could not complete Task 2 in the second session. Four of those subjects lost their sight in adulthood or late adolescence, reducing or eliminating the possibility of having been introduced to tactile maps in an educational setting. In addition, several of the subjects who did not complete Task 2, including those with congenital and adventitious vision loss, verbally indicated a lack of experience with tactile maps. All requested additional time to complete the tasks. In order to maintain consistency in the results, this request could not be fulfilled. While these subjects would likely never have completed as many tasks as their peers in the experiment, allotting greater time overall for all subjects could have enabled them to complete a

greater number of tasks. Hence, the time limitation reduced the amount of information which could be obtained from the study regarding the impact of the tactile map on these subjects' prosodics.

Another issue regarding time limitations on task completion merits discussion. For both the congenital and adventitious populations, certain variables differed significantly at the overall level and at the Task 1 level, but not for Task 2. Some subjects from both of these populations did not complete Task 2 in the multimodal session; thus, it was important to examine whether their behavior caused the differences exhibited for Task 1. While the statistical evidence showed some impact of these subjects on the differences exhibited for Task 1, it did not strongly support this hypothesis, and in one case, contradicted it. Without additional data, it is possible to attribute the differences exhibited for Task 1 to the discomfort for all subjects in adapting to the interface. Nonetheless, data from additional tasks beyond Task 2 could have offered further information on this issue.

Despite these limitations, however, this study produced a unique database of information for further investigation. Some final analyses of the data are presented in the next section. These pertain to the categorization of the sighted subjects by cognitive preference. This categorization was performed on the basis of subject responses to the Individual Differences Questionnaire (IDQ) detailed in Paivio and Harshman (1983).

Sighted Subjects Categorized by Cognitive Preference

All sighted subjects responded to the the IDQ, which seeks to elicit an individual's dominant cognitive style. The questionnaires were scored according to instructions in Paivio and Harshman (1983). Each subject was given a score on a verbal and a visual or "imaginal" scale. Of the 27 subjects whose data was used in the analyses, 21 subjects scored highest on the visual scale, while 6 subjects scored highest on the verbal scale. This produced two new populations, a "visual" population, consisting of 21 subjects, and a "verbal" population, consisting of 6 subjects. Thus, data for each of these populations was analyzed using a matched-pair t test to compare the means of the differences in the measurements of the prosodic features produced in the displayless sessions against the multimodal sessions; also, a matched-pair t test was performed comparing the number and type of recognition errors made by the system on utterances spoken by users in the displayless versus the multimodal sessions. The results of these analyses are given in Tables 79-84. To summarize the results, only those variables pertaining to pauses differed significantly between sessions for both populations. All other variables, including "L%" and "H%" boundary tones, duration of utterances, and number of recognition errors made by the system differed significantly only for the population scoring highest on the visual scale.

Table 79. Matched-pair T Test Results for Number of Hesitation Pauses per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff.	Std. Error	T	Prob(T)
Visual	1.333333	0.3507752	3.957477	0.0014 ***
Verbal	1.833333	0.6009252	3.0508511	0.0284 **

'***' Indicates difference was significant at alpha <= 0.01.
'**' Indicates difference was significant at alpha <= 0.05.

Table 80. Matched-pair T Test Results for Average Length of Hesitation Pauses per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff. (Sec)	Std. Error	T	Prob(T)
Visual	0.122477	0.0594991	2.054614	0.0532 *
Verbal	0.239745	0.0699176	3.428974	0.0187 **

'**' Indicates difference was significant at alpha <= 0.05.
'*' Indicates difference was significant at 0.05 <= alpha <= 0.06.

Table 81. Matched-pair T Test Results for Number of "L%" Boundary Tones per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff.	Std. Error	T	Prob(T)
Visual	13.3333	4.0803439	3.2676984	0.0039 ***
Verbal	15.166667	8.0266916	1.8895290	0.1174

'***' Indicates difference was significant at alpha <= 0.01.

Table 82. Matched-pair T Test Results for Number of "H%" Boundary Tones per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff.	Std. Error	T	Prob(T)
Visual	5.5	2.8707217	1.9241884	0.0587 *
Verbal	1.0	1.8257419	0.5477226	0.6074

'*' Indicates difference was significant at 0.05 <= alpha <= 0.06.

Table 83. Matched-pair T Test Results for Duration of Utterances Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff.	Std. Error	T	Prob(T)
Visual	0.1705914	0.0588862	2.8969672	0.0089 ***
Verbal	0.1158063	0.1638901	0.7066096	0.5114

'***' Indicates difference was significant at alpha <= 0.01.

Table 84. Matched-pair T Test Results for Number of Substitution Errors Made by the System per Utterance Spoken by Sighted Subjects in Displayless vs. Multimodal Sessions, Given by Subject Cognitive Preference

Subject Cognitive Preference	Mean Diff.	Std. Error	T	Prob(T)
Visual	3.89	1.0077928	3.8273175	0.0011 ***
Verbal	2.00	1.4142136	1.411236	0.2164

'***' Indicates difference was significant at alpha <= 0.01.

Clearly, the population scoring highest on the visual scale exhibited the greatest number of significant changes in prosodics between sessions. Also, the number of recognition errors made by the system was significantly greater for this population during the displayless sessions than the multimodal sessions. It is possible to infer from these results that this population experienced greater cognitive stress in the absence of the visual display than that of the population scoring higher on the verbal scale. The small sample size for the latter population, however, weakens such an inference. Though the larger sample size of those scoring higher visually is more likely to be representative of the population as a whole, the relatively small group of those scoring higher verbally, i.e., 6 subjects, is less likely to be so and more likely to be taken from the tails of the population. Nonetheless, the strength of the findings for those scoring higher visually warrants further examination of this issue in a follow-on study. If the low numbers of subjects scoring higher on the verbal scale in this experiment are

indicative of the larger population, a considerably larger overall sample size from the sighted population will be needed to conduct such a study.

Relevance of Results for Prosodic Pattern Detection Algorithms

All populations analyzed in this research exhibited significant differences for at least one prosodic feature when using the displayless navigational interface; for most populations, a combination of prosodic features differed significantly. These results strongly support the arguments of Wightman and Ostendorf (1994) that multiple prosodic features are needed for robust prosodic pattern detection algorithms.

Specific to this study, the universality of results concerning pauses indicates that this prosodic feature may not be the best predictor for phrase boundaries for displayless navigational applications. The differences in tonal and durational data, particularly for the sighted and adventitious populations, indicate that these features are also important for phrase boundary prediction algorithms. Further, the differences in boundary tones, particularly the significant increase in "L%" tones during displayless sessions, present difficulties for tune detection algorithms which seek to classify utterances as yes/no questions based on the ending tone in the utterance. Since significantly more utterances end in low declarative tones, it is more likely that a user may conclude a yes/no question in this manner, thus confounding algorithms which expect a high tone. Finally, similar problems arise for prominence detection algorithms which rely on a single acoustic cue, such as F0, to detect the speaker's emphasis. Given the variability in prosodic features during displayless sessions, a speaker is more likely to use a

combination of cues to indicate emphasis during these sessions, such as durational lengthening along with shifts in F0.

Since the database of speech produced from the experiments in this research was labeled prosodically by hand using the ToBI transcription system, many of these issues can be explored further. More generally, much of the work in prosodic pattern detection has relied on the use of either recorded speech read from a prepared text or from interactions with a speech surrogate. Very few databases of spontaneous speech with a live recognizer are available. Thus, the speech corpus produced from this research adds to the limited resources available for further investigation of the issues argued by Wightman and Ostendorf (1994).

CHAPTER VI

CONCLUSIONS

This research has investigated the hypothesis that the prosodic patterns of speech produced by users of displayless navigational interfaces differ significantly from those of speech produced when the interface employs an additional output modality. The hypothesis was tested through experiments in which sighted and visually impaired users employed a displayless and a multimodal speech-based prototype to perform a series of navigational tasks in an unfamiliar area. Recordings of user speech during these experiments were post-processed for prosodic content and statistically analyzed for differences between displayless and multimodal sessions. Samples from three populations were analyzed, including users with congenital vision loss, users with adventitious vision loss, and sighted users. These analyses were performed on overall session data as well as task-level data in order to observe the effect of increased spatial complexity of the tasks. While the results of analyses for each population differed in particular aspects, each exhibited significant differences in the prosodics of user speech during the displayless versus the multimodal sessions, both in overall comparisons and task-level comparisons.

Hesitation pauses were found to be significantly greater in number for all populations in displayless interface sessions. Other aspects of pauses differed among

populations. Recognition errors were also found to be significantly greater for all populations in displayless interface sessions, although the nature of the errors differed among populations. Beyond these general trends, results differed for each population; however, results for users with congenital vision loss were most dissimilar from the other populations, suggesting possible differences in the way in which these users adapt to this type of application. The hypothesis of this research assumed that prosodics reflect cognitive load. Therefore, according to that assumption, because this population exhibited fewest changes in prosodics during displayless sessions, they experienced least additional cognitive load.

Beyond revealing potential problems in the use of displayless navigational interfaces, this research provided a basis for improving their usability by gathering baseline observations of the factors which contributed to the increase in cognitive load. The most significant observation concerned the difficulty of users in maintaining a general sense of compass directions. Possible solutions to explore include augmenting the interface with either localized sound sources or a palm-sized visual or tactile map.

This investigation also produced results supporting previous work concerning prosodic pattern detection, specifically the use of multiple acoustic cues in prosodic pattern detection algorithms. Of equal importance, it contributed to a limited number of spontaneous speech databases available for prosodic pattern detection research.

Several areas for future investigation have been identified. As mentioned, one potential area for further study concerns the differences in the adaptive behavior of users with congenital vision loss and how these differences may relate to a lack of

visual memory. A possible strategy for examining this issue might include measuring the prosodics of this category of users when using a displayless application which is not spatially based, such as the airline reservation application used in research conducted by Zue et al. (1994). This could provide more complete information on the relationship on how the variables concerning the spatial nature of the application and the lack of visual memory affect cognitive load.

In addition to those previously mentioned, another area for exploration concerns the nature of deployment of the prototype. The prototype for the experiments in this research was developed and used in a stationary mode in an office environment, allowing users to plan routes before walking them. Deploying the prototype in an actual mobile environment with the noise and other distractions of a real situation could yield different results. This investigation offered the advantage of isolating, as much as possible, the spatial and verbal aspects of the navigational problem. Therefore, the results of this study compared to those of experiments conducted with a portable application would provide a richer source of knowledge than either alone.

Finally, the usage of the prototype, i.e., short-term or long-term, presents another area for investigation and could provide insight on issues pertaining to the dissimilarities in results for subjects with congenital vision loss. These users tended to more quickly seek out and find speech shortcuts for solving problems, thus exhibiting the behavior of an "expert" or longer-term user more quickly than sighted users or those with adventitious vision loss. The prototype used in this research was developed

for short-term or one-time users, i.e., visitors to the WES. However, many displayless applications target longer-term users, such as the personal guidance system developed by Loomis et al. (1994) or the Soldier's Computer (Weinstein 1994). An investigation comparing the behavior of short-term and long-term users of displayless applications could offer additional insight on user adaptation styles.

To conclude, displayless navigational technology offers users the possibility of a greater level of independence, whether as a visually impaired or sighted visitor in an unfamiliar area, or a soldier on the battlefield. This research has explored and elucidated some of the issues critical to its acceptance by the user community.

REFERENCES

Arias, C., C.A. Curet, H.F. Moyano, S. Joekes, N. Blanch. 1993. Echolocation: A study of auditory functioning in blind and sighted subjects. *Journal of Visual Impairment and Blindness* 87(3):73-77.

Baca, J., and K. Cooper. 1989. Modelling user goals and plans in natural language interfaces. In *Proceedings of the 27th Annual Southeast Regional ACM Conference, Atlanta, GA, April 1989*, 482-484.

Barth, J.L. 1983. Introducing blind students to reading tactile maps. *Chap. In Tactile Graphics Guidebook*, 27-33, Louisville, KY: American Printing House for the Blind.

Bear, J. and P. Price 1990. Prosody, syntax, and parsing. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, Pittsburgh, PA*, 17-22.

Beckman, M., and G. Ayers. 1997. *Guidelines for ToBI Labelling, version 3.0.* Manuscript and accompanying speech materials, Ohio State University.

Beckman, M. and J. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3:255-309.

Beggs, W.D.A. 1992. Coping, adjustment, and mobility-related feelings of newly visually impaired young adults. *Journal of Visual Impairment and Blindness* 86(3):136-139.

Berliss, J. 1993. *Non-Visual Human-Computer Interactions*, eds. D. Burger and J.C. Sperandio, 131-143, Montrouge, France: John-Libbey Eurotext.

Biermann, A.W., L. Fineman, and J.F. Heidlage. 1992. A voice and touch-driven natural language editor and its performance. *International Journal of Man-Machine Studies* 37:1-21.

Bigelow, A.E. 1986. The development of reaching in blind children. *British Journal of Developmental Psychology* 4:355-366.

Blattner, M., D. Sumikawa, and R. Greenburg. 1989. Earcons and icons: Their structure and common design principles. *Human Computer Interaction* 4(1): 11-44.

Bolinger, D. L. 1958. A theory of pitch accent in English. *Word* 14:109-149.

Bonner, M.R. 1943. Changes in the speech pattern under emotional tension. *The American Journal of Psychology* 56:262-273.

Boomer, D.S. 1968. Hesitation and grammatical and encoding. *Language and Speech* 8:148-158.

Bower, G.H., M.B. Karlin, and A. Dueck. 1975. Comprehension and memory for pictures. *Memory and Cognition* 3:216-220.

Boyd, L.H., W.L. Boyd, J. Berliss, M. Sutton, and G.C. Vanderheiden. 1992. The paradox of the graphical user interface: Unprecedented computer power for blind people. *Closing the Gap* 14 (October): 24-25, 60-61.

Boyd, L.H., W.L. Boyd, and G.C. Vanderheiden. 1990. The graphical user interface: crisis, danger, and opportunity. *Journal of Visual Impairment and Blindness* 84:496-502.

Bradford, J.H. 1995. The human factors of speech-based interfaces: A research agenda. *Association of Computing Machinery SIGCHI Bulletin* 27(2): 61-67.

Brenner, M., H.H. Branscomb, and G.E. Schwartz. 1979. Psychological stress evaluator: Two tests of a vocal measure. *Psychophysiology* 16:351-357.

Brieman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth and Brooks.

Brooks, L.R. 1968. Spatial and verbal components of the act of recall. *Canadian Journal of Psychology* 22:349-368.

Burger, J.D. and R.J. Marshall. 1993. The application of natural language models to intelligent multimedia. Chap. in *Intelligent Multimedia Interfaces*, ed. M. Maybury, 174-196, Menlo Park, CA: AAAI Press.

Burger, D. 1994. Improved access to computers for the visually handicapped: New prospects and principles. *IEEE Transactions on Rehabilitation Engineering* 2(3): 111-118.

Burger, D., C. Mazurier, S. Cesarano, and J. Sagot. 1993. *Non-Visual Human-Computer Interactions*, eds. D. Burger and J.C. Sperandio, 97-114, Montrouge, France: John-Libbey Eurotext.

Butcher, H.J. 1968. *Human Intelligence: Its Nature and Assessment*. New York: Harper and Row.

Butler, S.R. and A. Glass. 1974. Asymmetries in the electroencephalogram associated with cerebral dominance. *Electroencephalography and Clinical Neurophysiology* 36:48-491.

Butzberger, J.W., M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. 1990. Isolated word intonation recognition using Hidden Markov models. In *Proceedings International Conference Acoustics, Speech, and Signal Processing*, S-2:773-776.

Catell, R.B. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology* 54:1-22.

Catell, R.B. 1971. *Abilities: Their Structure, Growth, and Action*. Boston, MA: Houghton Mifflin.

Campbell, W.N. 1992. Prosodic encoding of speech. In *International Conference Spoken Language Processing, Banff, Canada*, 663-66.

Centigram. 1996. TruVoice Text-to-Speech Software Development Kit. Version 5.1.

Chen, F. and M. Withgott. 1992. The use of emphasis to automatically summarize a spoken discourse. In *Proceedings International Conference on Acoustics, Speech and Signal Processing* 1:229-232.

Chiarello, C. 1980. A house divided? Cognitive functioning with callosal agenesis. *Brain and Language* 11:128-158.

Clarke, S., R. Kraftsik, H. Van der Loos, and G.M. Innocenti. 1989. Forms and measures of adult and developing human corpus callosum: Is there sexual dimorphism? *Journal of Comparative Neurology* 280:213-230.

Cohen, P.R., M. Dalrymple, D.B. Moran, F.C.N. Pereira, J.W. Sullivan, R. Gargan, J. L. Schlossberg, S. W. Tyler. 1989. Synergistic use of direct manipulation and natural language. In *Proceedings of the CHI'89, held in Austin, Texas, April 30- May 4, 1989*, eds. K. Bice and C. Lewis, 227-245.

Coutaz, J., D. Salber, and S. Balbo. 1993. Towards automatic evaluation of multimodal user interfaces. *Knowledge-Based Systems* 6(4), 267-274.

Cronbach, L.J. 1957. The two disciplines of scientific psychology. *American Psychologist* 12:671-684.

Cronbach, L.J. and R.E. Snow. 1977. *Aptitudes and Instructional Methods: Handbook for Research on Interactions*. New York: Irvington.

Crudden, A. 1997. Congenital and adventitious vision loss: A comparison based on postemployment factors. Ph.D. diss., Mississippi State University.

Cruttenden, A. 1986. *Intonation*. Cambridge:Cambridge University Press.

Dale, B. 1992. Issues in Traumatic Blindness. *Journal of Visual Impairment and Blindness* 86(3):140-143.

Daly, N., and V. Zue. 1990. Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues. In *Proceedings International Conference Spoken Language Processing, Kobe, Japan*, 497-500.

da Lacoste-Utamsing, C. and R.L. Holloway. 1982. Sexual dimorphism in the human corpus callosum. *Science* 216:1431-1432.

Davidson, R.J., G.E. Schwartz, E. Pugash, and E. Bromfield. 1976. Sex differences in patterns of EEG asymmetry. *Biological Psychology* 4:119-138.

Davis, J.R., and C. Schmandt. 1989. The back seat driver: Real time spoken driving instructions. In *First Vehicle Navigation and Information Systems Conference, Toronto, Ontario, Canada, September 11-13, 1989*, 146-150.

Dodds, A.G., C.I. Howarth, and D.C. Carter. 1982. The mental maps of the blind: The role of previous visual experience. *Journal of Visual Impairment and Blindness* 76:5-12.

Dowdy, S. And S. Wearden. 1983. *Statistics for Research*. New York: John Wiley and Sons.

Drake, R.M. 1954. *Manual for Drake Musical Aptitude Tests*. Chicago: University of Chicago Press.

Duez, D. 1985. Perception of silent pauses in continuous speech. *Language and Speech* 28(4):377-389.

Duman, R. and S. Morgan. 1975. EEG asymmetry as a function of occupation, task and task difficulty. *Neuropsychologia* 13:219-228.

Edwards, W.K., E.D. Mynatt, and T. Rodriguez. 1993. The Mercator project: A nonvisual interface to the X Window System. *The X Resource*. Sebastopol, CA: O'Reilly and Associates, Inc.

Edwards, W.K., E.D. Mynatt, and K. Stockton. 1994. Providing access to graphical user interfaces - not graphical screens. In *The First Annual ACM Conference on Assistive Technologies, held in Los Angeles, CA, October 31-November 1, 1994*, 47-54.

Environmental Systems Research Institute. 1995. ARC/INFO. Version 7 for UNIX and OpenVMS. Redlands, CA: Environmental Systems Research Institute.

Entropic Research Laboratories. 1996. HTK The Hidden Markov Model Toolkit. Version 2.1. Cambridge, England: Entropic Cambridge Research Laboratory.

Ericsson, K.A. and A. Simon. 1993. Effects of verbalization. Chap. in *Protocol Analysis*, 63-106, Cambridge, MA: MIT Press.

Fairbanks, G. and W. Pronovost. 1939. An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs* 6:87-104.

Fairbanks, G. and L.W. Hoaglin. 1941. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs* 8:85-90.

Ferrell, K.A. 1986. Infancy and early childhood. In *Foundations of Education for Blind and Visually Handicapped and Youth*, ed., G.T. Scholl, New York: American Foundation for Blind.

Fraiberg, S. 1977. *Insights for the Blind*. New York: Basic Books.

Galin, D. and R. Ornstein. 1962. Lateral specialization of cognitive mode: An EEG study. *Psychophysiology* 9:412-418.

Gaver, W. 1989. The SonicFinder: An interface that uses auditory icons. *Human Computer Interaction* 4(1): 67-94.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech.* London: Academic Press.

Goldman-Eisler, F. 1972. Pauses, clauses, sentences. *Language and Speech* 15:103-113.

Griffin, H.C. 1981. Motor development in congenitally blind children. *Education of the Visually Handicapped* 7:107-111.

Grosjean, F., and M. Collins. 1979. Breathing, pausing and reading. *Phonetica* 36, 98-114.

Guilford, J.P. 1967. *The Nature of Human Intelligence*. New York: McGraw-Hill.

Halpern, D.F. 1987. *Sex Differences in Cognitive Abilities*. Hillsdale, NJ: Erlbaum.

Hart, R. and G. Moore. 1973. The development of spatial cognition: A review. Chap. in *Image and Environment: Cognitive Mapping and Spatial Behavior*, eds., R. Downs and D. Stea, 246-288. Chicago: Aldine.

Hendrix, G.G. 1978. Semantic aspects of translation. Chap. in *Understanding Spoken Language*, ed., D. Walker, 193-226. New York: Elsiver Science Publishing Co.

Hendrix, G.G., E.D. Sacerdoti, D. Sagalowicz, and J. Slocum. 1978. Developing a natural language interface to complex data. In *ACM Transactions on Database Systems*, 3(2):105-147.

Hieronymous, J., D. McKelvie, and F. McInnes. 1992. Use of acoustic sentence level and lexical stress in HSSM speech recognition. In *Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, held in San Francisco, CA, March 23-26, 1992*, 225-227.

Hill, D.R. and C. Grieb. 1988. Substitution for a restricted visual channel in multimodal computer-human dialogue. *IEEE Transactions on Systems, Man, and Cybernetics* 18(2):285-403.

Hill, E.W., and P.E. Ponder. 1976. *Orientation and Mobility: A Guide for the Practitioner*. New York: American Foundation for the Blind.

Hill, E.W., J.J. Riser, M.M. Hill, M. Hill, J. Halpin, and R. Halpin. 1993. How persons with visual impairments explore novel spaces: Strategies of good and poor performers. *Journal of Visual Impairment and Blindness* 87(8):295-301.

Hines, M., L.A. McAdams, L. Chiu, and P.M. Bentler. 1992. Cognition and the corpus callosum: Verbal fluency, visuospatial ability, and language lateralization related to midsagittal surface areas of callosal subregions. *Behavioral Neuroscience* 106(1):3-14.

Hinton, R.A.I. 1991. Use of tactile pictures to communicate the work of visual artists to blind people. *Journal of Visual Impairment and Blindness* 85(4):174-175.

Hixon, T.J., D.H. Klatt, and J. Mead. 1971. Influence of forced transglottal pressure change on voice fundamental frequency. *Journal of Acoustical Society of America* 49: 446-457.

Holden, C. 1975. Lie detectors: PSE gains audience despite critics' doubts. *Science* 190:359-362.

Honma, S., and R. Nakatsu. 1987. Dialogue analysis for continuous speech recognition. In *Record of the Annual Meeting of the Acoustical Society of Japan*, 105-106.

Huber, D. 1989. A statistical approach to the segmentation and broad classification of continuous speech into phrase-sized information units. In *Proceedings International Conference on Acoustics, Speech, and Signal Processing, held in Glasgow, Scotland, May 1989*, 600-603.

Hudson, D. 1994. Causes of emotional and psychological reactions to adventitious blindness. *Journal of Visual Impairment and Blindness* 88(10):498-503.

Jacobson, W.H. 1993. Basic outdoor O&M skills. Chap. in *The Art and Science of Teaching Orientation and Mobility to Persons with Visual Impairments*, 105-116, New York, NY: AFB Press.

Juurmaa, J. 1967. The ability structure of the blind and the deaf: Final report. *American Foundation for the Blind Research Bulletin* 14:109-122.

Kamm, Candace. 1994. User interfaces for voice applications. Chap. in *Voice Communication Between Humans and Machines*, eds. D.B. Roe and J.G. Wilpon, 422-444, Washington, D.C.: National Academy Press.

Kieras, D. 1978. Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin* 85:532-554.

Klatzky, R.L. 1980a. Visual codes in short term memory. Chap. in *Human Memory Structure and Processes*. New York: Freeman and Company.

Klatzky, R.L. 1980b. Representation of visual information in LTM. Chap. in *Human Memory Structure and Processes*. New York: Freeman and Company.

Kosslyn, S.M. and S.P. Schwartz. 1977. A simulation of visual imagery. *Cognitive Science* 1:265-296.

Kozlowski, L., and K. Bryant. 1977. Sense of direction, spatial orientation, and cognitive maps. *Journal of Experimental Psychology* 3(2):590-598.

Krause, J. 1993. A multilayered empirical approach to multimodality: Towards mixed solutions of natural language and graphical interfaces. Chap. in *Intelligent Multimedia Interfaces*, ed. M. Maybury, 307-327, Menlo Park, CA:AAAI Press.

Kuroda, I., O. Fujiwara, and N. Okamura. 1976. Method for determining pilot stress through analysis of voice communication. *Aviation, Space, and Environmental Medicine* 47:528-533.

Liberman, P. 1967. *Intonation, Perception and Language*. Cambridge, MA:MIT Press.

Ljolje, A. and F. Fallside. 1987. Recognition of isolated prosodic patterns using hidden Markov models. *Computer Speech and Language* 2:27-33.

Loomis, J.M., R.G. Golledge, R.L. Klatzky, J. Speigle, and J. Tietz. 1994. Personal guidance system for the visually impaired. In *ASSETS '94, The First Annual ACM Conference on Assistive Technologies, October 31-November 1, 1994, Los Angeles, CA*, 85-91.

Maccoby, E.E. and C.N. Jacklin. 1974. *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.

Marsh, E., K. Wauchope, and J.O. Gurney, Jr. 1994. Human-machine dialogue for multi-modal decision support systems. *AAAI Spring Symposium Series on Intelligent Multi-Modal Multi-media Systems, Stanford University, Palo Alto, CA*, (NCARAI Report. AIC-94-032).

McLinden, D.J. 1988. Spatial Task Performance: A meta-analysis. *Journal of Visual Impairment and Blindness* 82(6):231-236.

Moore, Robert. C. 1994. Integration of speech with natural language understanding. In *Voice Communication Between Humans and Machines*, eds. D.B. Roe and J.G. Wilpon, 255-269, Washington, D.C:National Academy Press.

Mulholland, T. 1978. A program for the EEG study of attention in visual communication. Chap. in *Visual Learning, Thinking, and Communication*, eds. B.S. Randhawa and W.E. Coffman, 77-91, New York: Academic Press.

Mynatt, E.D., and G. Weber. 1994. Nonvisual presentation of graphical user interfaces:contrasting two approaches. In *Proceedings of the ACM CHI '94 Conference, held in Boston, MA, April 24-28, 1994*, 166-172.

Nakai, M., H. Shimodaira, and S. Sagayma. 1994. Prosodic phrase segmentation based on pitch-pattern clustering. *Electronics and Communications in Japan* 77(6): 80-91.

Ngan, J., and J. Picone. 1997. Issues in generating pronunciation dictionaries for voice interfaces to spatial databases. In *Proceedings of the IEEE Southeastcon '97, Blacksburg, VA, April 12-14, 1997*, 97-99.

Ochaita, E., and J.A. Huertas. 1993. Spatial representation by persons who are blind: A study of the effects of learning and development. *Journal of Visual Impairment and Blindness* 87(2): 37-41.

Ohala, J. and W. Ewan. 1973. Speed of pitch range. *Journal of Acoustical Society of America* 53:345.

Okawa, S., T. Endo, T. Kobayashi, and K. Shirai. 1993. Phrase recognition in conversational speech using prosodic and phonemic information. *IEICE Transactions of Information and Systems* E76-D(1): 44-50.

Paivio, A. 1971. *Imagery and Verbal Processes*. New York: Holt, Rinehart, and Winston.

Paivio, A., and R. Harshman. 1983. Factor analysis of a questionnaire on imagery and verbal habits and skills. *Canadian Journal of Psychology* 37(4):461-483.

Pierrehumbert, J. and J. Hirschberg. 1990. The meaning of intonation contours in the interpretation of discourse. Chap. in *Intentions in Communication*, eds., P.R. Cohen, J. Morgan, and M.E. Pollack, Cambridge, MA:MIT Press.

Pitman, D.J. 1965. The musical ability of blind children. *American Foundation for the Blind Research Bulletin* 11: 673-679.

Price, P., M. Ostendorf, S. Shattuck-Hufnagel, C. Fong. 1991. The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America* 90:2956-2970.

Pylyshyn, Z.W. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* 80:1-24.

Rafnel, K.J. and R.L. Klatzky. 1978. Meaningful-interpretation effects on codes of nonsense pictures. *Journal of Experimental Psychology: Human Learning and Memory* 4:631-646.

Resnick, R. 1983. An exploratory study of the lifestyles of congenitally blind adults. *Journal of Visual Impairment and Blindness* 77(10):476-481.

Rice, C.E. 1969. Perceptual enhancement in the early blind? *Psychological Record* 19:1-14.

Rice, C.E. 1970. Early blindness, early experience and perceptual enhancement. *American Foundation for the Blind Research Bulletin* 22:1-22.

Robertson, I.T. 1985. Human information-processing strategies and styles. *Behaviour and Information Technology* 4(1):19-29.

Rochester, S.R. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research* 2(1):51-81

Rosenfeld, R., Byrne, B., Iyer, R., Liberman, M., Shriberg, L., Unveferth, J., Vidal, E., Agarwal, R., and D. Vergyri. 1996. Error analysis and disfluency modeling in the Switchboard domain. In *Proceedings of the International Conference on Spoken Language Processing, 1996, Philadelpha, PA*, SAP1S1.3.

Scherer, K.R. 1981. Speech and emotional states. Chap. in *Speech Evaluation in Psychiatry*, ed., J.K. Darby, 189-220, New York: Grune-Stratton.

Schmandt, C. 1994. Using speech recognition. Chap. in *Voice Communication with Computers, Conversational Systems*, 154-178, New York: Van Nostrand Reinhold.

Schuck, J.R. 1973. Factors affecting reports of fragmenting visual images. *Perception and Psychophysics* 18: 382-390.

Schuck, J.R. and W.R. Leahy. 1966. A comparison of verbal and non-verbal reports of fragmenting visual images. *Perception and Psychophysics* 1:191-192.

Shepard, R.N. 1967. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior* 6:156-165.

Shimodaira, H., and M. Kimura. 1992. Accent phrase segmentation using pitch pattern clustering. In *Proceedings International Conference on Acoustics, Speech, and Signal Processing, held in San Francisco, CA, March 23-26, 1992*, 217-220.

Shneiderman, B. 1984. Direct manipulation. Chap. in *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 200-229, Don Mills Ontario: Addison-Wesley Publishing Company.

Shneiderman, B. 1992. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Second Edition, 280-281, Don Mills Ontario:Addison-Wesley Publishing Company.

Silverman, K.E.A., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992a. TOBI: A standard for labelling English prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICLSP), Banff Alberta, Canada, October 1992*, 867-870.

Silverman, K., E. Blaauw, J. Spitz, and J.F. Pitrelli. 1992b. A prosodic comparison of spontaneous speech and read speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICLSP), Banff Alberta, Canada, October 1992*, 880-891.

Snow, R.E. 1977. Individual differences, instructional theory, and instructional design. Technical Report No. 4, Aptitude Research Project, School of Education, Stanford University.

Sperry, R. 1982. Some effects of disconnecting the cerebral hemispheres. *Science* 217:1223-1226.

Standing, L., J. Conezio, and R.N. Haber. 1970. Perception and memory for pictures: Single-trial learning of 2560 visual stimuli. *Psychonomic Science* 19:73-74.

Stankov, L. and G. Spilsburg. 1978. The measurement of auditory abilities of blind, partially sighted and sighted children. *Applied Psychological Measurements* 2:491-503.

Stifelman, L., B. Arons, C. Schmandt, and E. Hulteen. 1993. VoiceNotes: A speech interface for a hand-held voice notetaker. In *Proceedings of ACM InterCHI'93 Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, April 24-29, 1993*, 179-186.

Stock, O. 1994. Natural language in multimodal human-computer interfaces. *IEEE Expert*, April 1994, 40-44.

Streeter, L.A., D. Vitello, and S.A. Wonsiewicz. 1985. How to tell people where to go: Comparing navigational aids. *International Journal of Man/Machine Systems* 22(5): 549-562.

Thorndyke, P., and C. Stasz. 1980. Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology* 12:137-175.

Thorsen, N. 1980. A study of the perception of sentence intonation--Evidence from Danish. *Journal of Acoustical Society of America* 67(3):1014-1030.

Uslan, M. Schreier, E., and A. Meyers. 1990. A quick look at the NOMAD, an audio-tactile graphics processor. *Journal of Visual Impairment and Blindness* 84: 383-384.

Vanderheiden, G.C., and D.C. Kunz. 1990. Systems 3: An interface to graphic computers for blind users. In *Proceedings of the 13th Annual Conference of RESNA in Washington, D.C., June 20-24, 1990*, 150-200.

Vaissiere, J. 1983. Language-independent prosodic features. Chap. in *Prosody: Models and Measurements*, eds., A. Cutler and D.R. Ladd, 53-66, Berlin: Springer-Verlag.

Vernon, P.E. 1950. *The Structure of Human Abilities*. London: Methuen.

Waibel, A. 1988. *Prosody and Speech Recognition*. San Mateo, CA: Morgan Kaufmann.

Wang, M.Q., and J. Hirschberg. 1991. Predicting intonational boundaries automatically from text: the ATIS Domain. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, CA, February 19-22, 1991*, Sponsored by Defense Advanced Research Projects Agency Information Science and Technology Office, 378-381.

Weinstein, C.J. 1994. Military and government applications of human-machine communication by voice. Chap. in *Voice Communication Between Humans and Machines*, eds. D.B. Roe and J.G. Wilpon, 357-370, Washington, D.C.: National Academy Press.

Welsh, R.L., and D.W. Tuttle. 1997. Congenital and adventitious blindness. Chap. in *Foundations of Rehabilitation Counseling with Persons Who Are Blind or Visually Impaired*, eds., J.E. Moore, W. Graves, and J.B. Patterson, 60-79. New York: AFB Press.

Wightman, C.W. and M. Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing* 2(4):469-481.

Wightman, C.W., N.M. Veilleux, and M. Ostendorf. 1991. Use of prosody in syntactic disambiguation: An analysis-by-synthesis approach. In *Speech and Natural Language, Proceedings of a Workshop held at Pacific Grove, CA, February 19-22, 1991*, Sponsored by Defense Advanced Research Projects Agency Information Science and Technology Office, 384-389.

Williams, C.E. and K.N. Stevens. 1981. Vocal correlates of emotional states. Chap. in *Speech Evaluation in Psychiatry*, ed. J.K. Darby, 221-239, New York:Grune-Stratton.

Williams, C.E. and K.N. Stevens. 1972. Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* 4(2):1238-1250.

Williams, C.E. and K.N. Stevens. 1969. On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine* 40:1369-1372.

Witelson, S. 1976. Sex and the single hemisphere: Specialization of the right hemisphere for spatial processing. *Science* 191:425-427.

Yalow, E. 1980. Individual differences in learning from verbal and figural materials. School of Education, Stanford University, Aptitudes Research Project Report No. 12, Palo Alto, CA.

Yankelovich, N. G. Levow, and M. Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *Proceedings of CHI '95 Conference on Human Factors in Computing Systems, Denver CO, May 7-11, 1995*, 369-376.

Zoltan-Ford, E. 1991. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies* 34:527-547.

Zue, V., S. Seneff, J. Polifroni, M. Philips, C. Pao, D. Goddeau, J. Glass, and E. Bill. 1994. PEGASUS: A spoken language interface for on-line air travel planning. In *Proceedings of the Human Language Technology Workshop, Plainsboro, NJ*, 201-206.

APPENDIX A

SUBJECT INSTRUCTIONS FOR EXPERIMENT

The Waterways Experiment Station, or WES (pronounced Wes), is a U.S. Army Corps of Engineers research facility, located in Vicksburg, MS. It was established in response to the Mississippi River Flood of 1927. WES's role as the first federal hydraulics research facility was to help the Mississippi River Commission develop and implement a flood control plan for the lower Mississippi Valley. From these beginnings, WES has grown to become a large research facility.

Today there are over 1400 employees. Over 700 of these employees are engineers and scientists who are widely known and respected for their work in such diverse areas as hydraulics, oceanography, electronics, computer science, ecology, and environmental engineering, among numerous other scientific and engineering disciplines. WES research is carried out in six separate, but closely interrelated, laboratories: Hydraulics, Coastal Engineering, Geotechnical, Structures, Environmental, and Information Technology. These laboratories are housed in separate buildings which spread across an area of approximately 750 acres. The main entrance to the station is located two miles south of interstate 20 on Halls Ferry Road. Halls Ferry Road runs along the western boundary of the station. Approximately two miles east of Halls Ferry, Porters Chapel Road runs along the eastern boundary. Landmarks along the north/south boundaries include the Information Technology Laboratory near the northern boundary and the Structures Laboratory near the southern boundary. Due to its unique historical landmarks, such as a large-scale physical model of the Niagara Falls, and its lush natural setting , the WES entertains many visitors each year on a daily basis.

In these experiments, you will be asked to play the role of a first-time visitor to the WES. You will be given a series of navigational tasks to perform which will require getting from one location at the WES to another. More specifically, for each task you will be given two locations: the first will be your starting point and the second will be your destination. For example, your starting point might be the Environmental Laboratory and your destination, the Hydraulics Laboratory. You must determine how to get from your starting point to your destination. To help you accomplish your tasks, you will be using an automated voice-activated navigation system, called WES Auto Travel. The system will plan a driving route through the WES and give you spoken directions along the route, one segment at a time. There is one important condition of your task, however: You must assume that you will be walking from your starting point to your destination. Since WES Auto Travel can only guarantee the quality of its driving routes, you must check each segment of the route and decide if it is passable by you as a pedestrian. Fortunately, the navigational system knows certain facts about the conditions of each road and road segment that can be helpful to you in making your decision. Some of these conditions include whether or not the road has a sidewalk, the width of the shoulder, as well as the traffic patterns and speeds of motor vehicles on the road. Other attributes of the road known by WES Auto Travel include physical landmarks, some of which may help you decide if the road is passable by a pedestrian. These landmarks include things you would pass along the road, such as deep ravines or

drop-offs near the shoulder, ponds, lakes, bridges, guardrails, steep hills or sharp curves in the road, and buildings which may be surrounded by sidewalk even if the road itself does not have adjacent sidewalk. The program will provide some prompts explaining what it knows as well as help on how to ask for what it knows. You will be given four navigational tasks to perform and your speech will be recorded for each task and later analyzed. Do you have any questions?

APPENDIX B

SUBJECT QUESTIONNAIRES

Questionnaire for Subjects with Sight Loss

Personal Information

1. Name: _____

2. Gender: 2. _____
   Male (0) Female (1)

3. How old are you? 3. _____

4. What is the highest grade of school you actually completed? 4. a. _____
   (e.g., 1 year of college equals 13, etc.)

   b. (High School Graduates: Did you receive a certificate (1) b. _____
   or diploma (0) for high school?)

5. Occupation and/or Job Title: _____

6. How would you describe your vision? 6. _____

   (1) No usable vision

   (2) Very little usable vision

   (3) Quite a bit of usable vision

7. Results of most recent eye exam/low vision exam 7. a. _____
   a. Date:

   b. Primary Visual Diagnosis: _____

   c. Near Acuity with Best Correction: _____

   d. Distance Acuity with Best Correction: _____

   e. Visual Field Problems: (1) Yes (2) No 7. e. _____

   f. Contrast Sensitivity Problems: (1) Yes (2) No 7. f. _____

8. How old were you when you became legally blind, or were you born blind? (Indicate age in years or enter 00 if born blind) 8. _____

Lifestyle Information

9. Do you use Braille regularly? 9. a. _____

a. (1) Yes (2) No

b. If yes, do you use Grade I or Grade II Braille? 9. b. _____
(1) Grade I
(2) Grade II

10. Do you use a long cane? 10. _____

(1) Yes (2) No

11. Do you use a guide dog? 11. _____
(1) Yes (2) No

12. Have you received any orientation and mobility training services?

a. (1) Yes (2) No 12. a. _____

b. If yes, indicate length in days. 12. b. _____

13. How would you describe your O&M skills? 13. _____

(5) Excellent
(4) Above Average
(3) Average
(2) Below Average
(1) Poor

14. Do you have any other disabilities or health problems, e.g., diabetes, hypertension, etc.?:

_____

_____

15. How do you prefer to read? 15. _____

Regular print (1)
Large print (2)
Tape cassette/Talking book (3)
Computer disk (4)
Braille (5)
Other (6) _____

16. How long have you used computers? (Record in months.)

16. _____

17. Do you use a screen magnification program for computer usage?

17. _____

(1) Yes (2) No

18. Do you use synthesized speech for computer usage?

18. _____

(1) Yes (2) No

19. If you answered yes to both of the above, how often do you use each, e.g., 50% for each, or 20% screen magnification, 80% synthetic speech, etc.?

a. Percentage for screen magnification
b. Percentage for synthesized speech

19. a. _____

b. _____

20. Do you use a refreshable Braille display? 20. a. _____

a. (1) Yes (2) No

b. If you answered yes to the above, how often do you use it?

20. b. _____

(1) Not very often
(2) Only for certain applications
(3) Quite a bit

Questionnaire for Subjects without Sight Loss

Part I: Personal Information

1. Name: _____

2. Gender: 2. _____
   Male (0) Female (1)

3. How Old Are You? 3. _____

4. What is the highest grade of school you actually completed? 4.a._____
   (e.g. 1 year of college equals 13, etc.)

   b. (High School Graduates: Did you receive a certificate (1) b._____
   or diploma (0) for high school?)

5. Occupation and/or Job Title: _____

Part II. See next page for Cognitive Preference Questionnaire.

Cognitive Preference Questionnaire

Instructions: The statements in this questionnaire represent ways of thinking, studying and problem solving, which are true for some people and not for others. Read each statement and decide whether or not it is true with respect to yourself. Then indicate your answer in the column to the right. If you agree with the statement or decide that it does describe you, answer TRUE. If you disagree with the statement or feel that it is not descriptive of you, answer FALSE. Answer the statements as carefully and honestly as you can. The statements are not designed to assess the goodness or badness of any way you think. They are attempts to discover the methods of thinking you consistently use in various situations. There are no right or wrong answers. Answer every statement either true or false, even if you are not completely sure of your answer.

TRUE/FALSE

I often have difficulty explaining things to others. _____

I can usually express my thoughts clearly. _____

I often have ideas that I have trouble expressing in words _____

I have no difficulty in expressing myself verbally. _____

I have difficulty expressing myself in writing. _____

I am fluent at writing essays and reports. _____

I can easily think of synonyms for words. _____

I find it difficult to find enough synonyms or alternate forms of words when writing. _____

Essay writing is difficult for me. _____

I have better than average fluency in using words. _____

I am a good story teller. _____

I enjoy doing work that requires the use of words. _____

I tell jokes and stories poorer than most people. _____

I have difficulty producing associations for words. _____

I am usually able to say what I mean in my first draft of an essay or letter. _____

I have a large vocabulary. _____

My knowledge and use of grammar needs much improvement. _____

I would rather work with ideas than words. _____

I am good at thinking up puns. _____

My vocabulary is not as large as I would like. _____

If given the choice, I would rather listen to a good speaker than visit an art gallery. _____

I often use mental images or pictures to help me remember things. _____

My thinking often consists of mental pictures or images. _____

I find it difficult to form a mental picture of anything. _____

When remembering a scene, I use verbal descriptions rather than mental pictures. _____

I never use mental pictures or images when trying to solve problems. _____

I often enjoy the use of mental pictures to reminisce. _____

I can close my eyes and easily picture a scene I have experienced. _____

I think that most people think in terms of mental pictures whether they are completely aware of it or not. _____

I can easily picture moving objects in my mind. _____

I do not form a mental picture of people or places when reading of them. _____

When someone describes something that happens to him, _____

I sometimes find myself vividly imagining the events that happened.

I have only vague visual impressions of scenes I have experienced. _____

Listening to someone recount his experiences does not usually arouse mental pictures of the incidents being described. _____

I don't believe that anyone can think in terms of mental pictures. _____

When reading fiction I usually form a mental picture of a scene or room that has been described. _____

When doing mental arithmetic, such as addition, I think in abstract terms rather than actually picturing the numbers. _____

While I have often seen pictures of him, I cannot remember exactly what President Clinton looks like. _____

I often remember work I have studied by imagining the page on which it is written. _____

I would rather have a verbal description of an object or person, than a picture. _____

I find it easy to visualize the faces of people I know. _____

I can add numbers by imagining them to be written on a blackboard. _____

I can easily form a mental picture of Vice President Gore. _____

I prefer to read instructions about how to do something, rather than have someone show me. _____

I cannot generate a picture of a friend's face when I close my eyes. _____

I feel a picture is worth a thousand words. _____

I think that puns are the lowest form of humor. _____

It bothers me when I see a word used improperly. _____

I take great pains to express myself with precision and accuracy in both verbal speech and written work. _____

I am continually aware of sentence structure. _____

Studying the use and meaning of words has become a habit with me. _____

I spend very little time attempting to increase my vocabulary. _____

I speak or write what comes into my head without worrying greatly about my choice of words. _____

I enjoy learning new words and incorporating them into my vocabulary. _____

I enjoy being able to rephrase my thoughts in many ways for variety's sake when both writing and speaking. _____

I am disturbed by people who quibble about word usage. _____

The proper use of words is secondary to the ideas and content of speech or writing. _____

When I hear or read a word, a stream of other words often comes to mind. _____

I find that I am more critical of writing style than content when reading literature. _____

I have found it easy in the past to learn a second language. _____

Not enough people pay attention to the manner in which they express themselves. _____

I enjoy solving crossword puzzles and other games. _____

I read rather slowly. _____

I consider myself a fast reader. _____

I read a great deal. _____

My grades have been hampered by inefficient reading. _____

I enjoy visual arts, such as paintings, more than reading. _____

I remember things I have done myself, much better than things I have read. _____

I find it easier to learn from a demonstration than from written instructions. _____

I enjoy reading an interesting story even if it is not particularly well written. _____

INSTITUTIONAL REVIEW BOARD APPROVAL FORM
FOR THE PROTECTION OF HUMAN SUBJECTS IN RESEARCH
MISSISSIPPI STATE UNIVERSITY

STATEMENT OF BOARD: **IRB DOCKET #** 96-208

This is to certify that the research proposal entitled Displayless Interface Access to Spatial Data: Effects on Speaker Prosodics

and submitted by: Name: Julie Baca

Department: Computer Science

Name of Advisor: Julia Hodges

to **Sponsored Programs Administration** for consideration has been reviewed by the Regulatory Compliance Officer or the IRB and approved with respect to the study of human subjects as appropriately protecting the rights and welfare of the individuals involved, employing appropriate methods of securing informed consent from these individuals and not involving undue risk in the light of potential benefits to be derived therefrom.

Administrative Approval Date: December 18, 1996

_____(A) Contingent upon receipt of _____

_____

__X__(B) All necessary documents were received.

Expedited Approval Date: _____

_____(A) Contingent upon receipt of _____

_____

_____(B) All necessary documents were received.

Full Board Approval Date: _____

_____(A) Contingent upon receipt of _____

_____

_____(B) All necessary documents were received.

Robyn B. Remotigue December 18, 1996

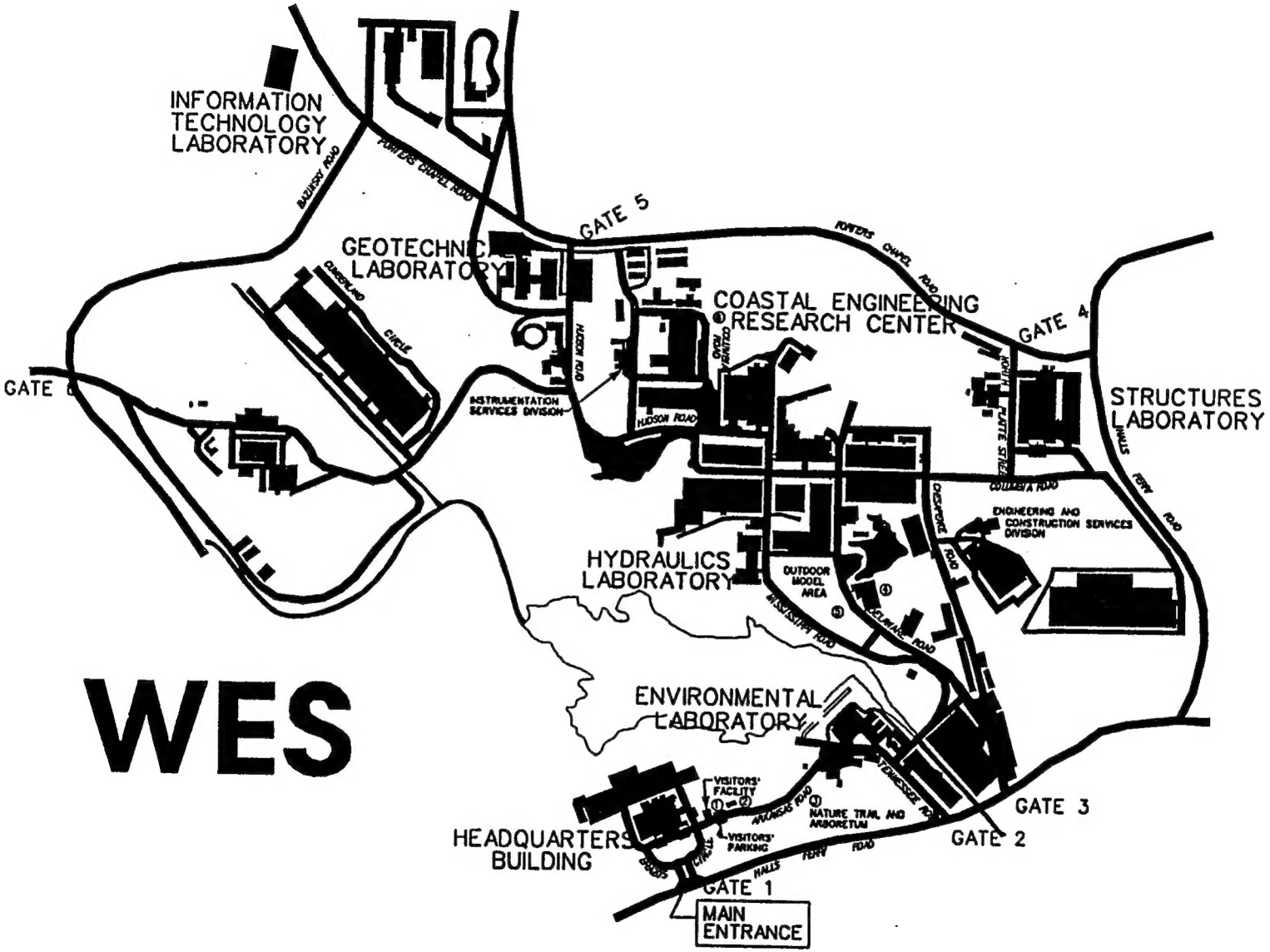
Robyn B. Remotigue, MSU Regulatory Compliance Officer **Date**

Institutional Review Board Member **Date**

(revised form 8/96)

APPENDIX C

MAP FOR MULTIMODAL SESSION


WES
INFORMATION TECHNOLOGY LABORATORY
GEOTECHNICAL LABORATORY
COASTAL ENGINEERING RESEARCH CENTER
HYDRAULICS LABORATORY
ENVIRONMENTAL LABORATORY
STRUCTURES LABORATORY
HEADQUARTERS BUILDING
ENGINEERING AND CONSTRUCTION SERVICES DIVISION
INSTRUMENTATION SERVICES DIVISION
OUTDOOR MODEL AREA
VISITORS' FACILITY
VISITORS' PARKING
NATURE TRAIL AND ARBORETUM
GATE 1
MAIN ENTRANCE
GATE 2
GATE 3
GATE 4
GATE 5
GATE 6
HUDSON ROAD
ARKANSAS ROAD
TENNESSEE ROAD
HALLS FERRY ROAD
CHEROKEE ROAD
COLUMBIA ROAD
FORT PLATTE STREET
LEWIS FERRY ROAD
CUMBERLAND CIRCLE
BAZINSKI ROAD
POINTERS CHAPEL ROAD

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Final report

4. TITLE AND SUBTITLE
Displayless Interface Access to Spatial Data: Effects on Speaker Prosodics

5. FUNDING NUMBERS

6. AUTHOR(S)
Julia A. Baca

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Waterways Experiment Station
3909 Halls Ferry Road, Vicksburg, MS 39180-6199

8. PERFORMING ORGANIZATION REPORT NUMBER
Technical Report ITL-98-3

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

Displayless interface technology provides speech-based access to computer programs for which visual access is not possible. These applications are increasingly prevalent, especially in situations requiring mobility, such as navigational applications, both civilian and military. To ensure the successful deployment of this technology, however, many human factor issues must be addressed. In particular, the nonvisual nature of this technology requires that it address a problem common to that of providing graphical user interface access to users with visual impairments, i.e., verbal presentation of spatial data. This research investigated a hypothesis rooted in the assumption that strictly verbal access to spatial data places a cognitive burden on the user. The prosodics, or nonverbal aspects, of human speech have been established in the literature as an indicator of cognitive stress. Therefore, this research examined the hypothesis that the cognitive burden placed on the user by displayless access to spatial data would impact the prosodics of the user's speech. Although the hypothesis was assumed to apply to all users, regardless of visual capability, differences in the manifestation of the impact on users with visual impairments versus sighted users were anticipated. Thus, both subjects with and without visual impairments participated in the research.

(Continued)

14. SUBJECT TERMS
Displayless
GUI
Prosodics

15. NUMBER OF PAGES
249

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED		

13. (Concluded).

The hypothesis was tested by conducting experiments in which user speech interactions with a prototype speech-based navigational system were recorded, postprocessed, and analyzed for prosodic content. Subjects participated in two sessions, one using a speech-based, displayless interface, and a second using a multimodal interface that included either a visual or tactile display. Subjects with visual impairments included both persons with adventitious as well as congenital sight loss. Results showed strong evidence of changes in subjects' prosodic features when using a displayless versus a multimodal navigational interface for all categories of subjects.